Generating Better Items for Cognitive Assessments Using Large Language Models

Antonio Laverghetta Jr. and **John Licato**
University of South Florida
Department of Computer Science and Engineering
Tampa, FL, USA
{alaverghett,licato}@usf.edu

Abstract

Writing high-quality test questions (items) is critical to building educational measures but has traditionally also been a time-consuming process. One promising avenue for alleviating this is automated item generation, whereby methods from artificial intelligence (AI) are used to generate new items with minimal human intervention. Researchers have explored using large language models (LLMs) to generate new items with equivalent psychometric properties to human-written ones. But can LLMs generate items with *improved* psychometric properties, even when existing items have poor validity evidence? We investigate this using items from a natural language inference (NLI) dataset. We develop a novel prompting strategy based on selecting items with both the best and worst properties to use in the prompt and use GPT-3 to generate new NLI items. We find that the GPT-3 items show improved psychometric properties in many cases, whilst also possessing good content, convergent and discriminant validity evidence. Collectively, our results demonstrate the potential of employing LLMs to ease the item development process and suggest that the careful use of prompting may allow for iterative improvement of item quality.

1 Introduction

AI is having increasingly profound impacts on educational and psychological measurement (Chen et al., 2020; Tavast et al., 2022). Technologies built on AI and machine learning, including educational data mining (Romero and Ventura, 2020), intelligent tutoring systems (Mousavinasab et al., 2021), deep item response theory (Cheng et al., 2019), and deep knowledge tracing (Piech et al., 2015), among others (Asfahani, 2022, inter-alia) are transforming educational and psychological measurement, and this trend seems likely to continue.

One promising educational application of large language models (LLMs) is for the automatic generation of test items (AIG). Writing high-quality test items is critical to building effective educational assessments, but has also traditionally been a time-consuming process, as items must be developed by experts and undergo numerous rounds of review (Bandalos, 2018). There has been significant research interest in using AIG to create high-quality items with minimal intervention to speed up the test development process (Prasetyo et al., 2020). Prior work has demonstrated that LLMs can generate items with at least face validity (i.e, they *appear* valid based on item content) for both non-cognitive (Götz et al., 2023) and cognitive (Attali et al., 2022) constructs. Careful psychometric analysis of items generated from such models has also revealed that they are just as valid and reliable as their human written counterparts (Lee et al., 2023). Although promising, this research has largely focused on generating items for constructs that have been well-studied, using items already known to have strong validity evidence. Suppose an educator wishes to develop a test for a new construct where existing items may have only undergone pretesting. Or suppose the educator wishes to use a new type of item for a well-established domain (e.g, a test of algebraic reasoning that uses a novel item format). In either case, the items will likely have limited validity evidence, and much time would need to be spent revising the items to improve their psychometric properties before they can be used.

In this work, we ask: can LLMs be used to generate valid and reliable items even in these scenarios where existing items have only limited validity evidence? If so, LLM-based AIG could be used to iteratively improve the psychometric properties of items, explore the underlying construct space, and shed light on what makes a good item.

We explore this using GPT-3 (Brown et al., 2020) and focus on generating items that test for natural language inference (NLI) (Dagan et al., 2006; Bowman et al., 2015). NLI is an important cognitive

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 414–428

construct in NLP research which, to our knowledge, has only undergone limited psychometric analysis in human participants (Laverghetta Jr. et al., 2021). We develop a novel prompting strategy that uses the psychometric properties of items, calculated using prior human responses, to select the most informative examples to send to the model to maximize the quality of the generated examples. **Our main contributions are as follows:**

1. We develop a novel prompting strategy for generating items by selecting items to include as context based on the psychometric properties they possess, focusing primarily on item discrimination.

2. Using GPT-3 we test our approach using the GLUE broad coverage diagnostic (Wang et al., 2018), a popular cognitive task in NLP research. We perform an extensive analysis of the psychometric properties of the generated items and find that those from GPT-3 show stronger evidence for validity and reliability than those written by humans in most cases.

2 Related Work

2.1 Automated Item Generation

Psychometricians have explored how to automate item generation for decades (Prasetyo et al., 2020). Early attempts focused on developing item models, which are systems that can interchange certain keywords in the item while keeping other parts of it constant (Bejar et al., 2002). While item models are theoretically justified and very likely to produce psychometrically valid items, developing them requires a great deal of manual effort, as both the item stem and other components must still be manually written. Furthermore, item models are limited in the diversity of content they can generate. These drawbacks have motivated recent work to investigate using LLMs as the item generator. von Davier (2018) was one of the first to explore this and used recurrent neural networks to generate items for a personality assessment. The advent of the transformer (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020) led to the creation of LLMs which could generate much more coherent and semantically accurate text, leading to further interest in LLM-based AIG. Götz et al. (2023) generated a large number of personality items using GPT-2 (Radford et al., 2019), and showed that at least some of these items passed face validity checks. Maertens et al. (2021) developed a test for misinformation susceptibility, using LLM-generated items. Hernandez and Nie (2022) developed a system for the automatic generation and validation of test items, using autoregressive LLMs for generation and autoencoding LLMs for validation. Lee et al. (2023) extensively evaluated the psychometric properties of GPT-3 generated personality items, including analysis of internal structure, differential item functioning, and reliability. They concluded that the validity evidence for machine-generated items was just as strong, if not stronger than, for human-written ones. While much work has focused on non-cognitive assessments, others have explored LLM-based AIG for educational assessments. Notably, Chan et al. (2022) used the BERT (Devlin et al., 2019) LLM to generate grammar reading exercises. Zou et al. (2022) and Rathod et al. (2022) used transformers to generate true/-false and reading comprehension questions. Attali et al. (2022) used transformer-based LLMs to generate items for the Duolingo English Test. Zu et al. (2023) used a combination of finetuning and prompt-based learning to train GPT-2 to generate distractors for fill-in-the-blank vocabulary items. A common theme throughout these works is the focus on well-studied assessments, and the use of items that have already been psychometrically validated in the prompt. Their goal is thus to generate items that maintain existing psychometric properties, which is different from our goal of generating items with *improved* properties.

2.2 Synthetic Data Generation in NLP

When it comes to gathering high-quality data, NLP researchers have concerns that overlap with those faced by the measurement community. Training examples for popular NLP tasks, including NLI (Bowman et al., 2015), and question answering (QA) (Rajpurkar et al., 2016), have historically been created using crowd-sourced annotations, which is both expensive and time-consuming. The incredibly rapid progress of LLMs in recent years also means that many once challenging datasets quickly become outdated as new models are developed (Ott et al., 2022). There has been significant research interest in using LLMs to generate synthetic training data, forgoing the need to run annotation studies (Schick and Schütze, 2021). Prior work has explored LLM-based data augmentation for QA (Duan et al., 2017), paraphrase identification

(Nighojkar and Licato, 2021), and NLI (Liu et al., 2022). Typically, this line of research relies on information-theoretic metrics of item quality, for example, dataset maps (Swayamdipta et al., 2020) to evaluate the newly generated items. Most relevant to our work is the study by Liu et al. (2022), who developed a system for using GPT-3 to automatically generate NLI items. However, their approach does not employ methods of assessing validity and reliability commonly used in educational measurement and instead relies on information-theoretic measures of item quality. Our goal is to generate items with improved validity and reliability in both human and LLM populations, using the psychometric properties of the items as the optimization target.

3 Generation of Test Items

The General Language Understanding Evaluation (GLUE) (Wang et al., 2018) is a benchmark designed to measure broad linguistic constructs in LLMs. Included in GLUE is a diagnostic set, AX,[1] which is meant to be a challenge set for diagnosing faults in LLMs. Items on AX are framed as NLI: given a premise (p) and hypothesis (h), a model must determine whether p entails, contradicts, or is neutral with respect to h (Dagan et al., 2006; Bowman et al., 2015). Items were written by NLP experts, inspired by categories taken from the FraCas suite (Cooper et al., 1996), and are based on sentences from a variety of artificial and naturalistic contexts. Wang et al. (2018) reported strong inter-rater reliability when labeling a random sample of AX items, and AX has been used successfully to evaluate many new LLMs (Brown et al., 2020; Raffel et al., 2020; Chowdhery et al., 2022), which suggests the diagnostic has good predictive validity. Furthermore, Laverghetta Jr. et al. (2021) previously ran human studies on a subset of items from AX, targeting those testing for propositional structure (PS), quantifiers (Q), morphological negation (MN), and lexical entailment (LE). Table 1 shows example AX items from these categories. They found that LLMs strongly predicted item difficulties and inter-item correlations in human responses across these categories, indicating good convergent validity for AX as a test of reasoning in both populations. Collectively, these results demonstrate a surface level of validity for the AX items (i.e, *face validity*); the items appear to function well in preliminary human studies and have been used successfully to find faults within LLM reasoning, but extensive analysis of their psychometric properties has yet to be performed. This makes AX a good assessment to use for our experiments, as we want items that have *not* undergone extensive psychometric development, and hence may not have strong validity as measures of the construct in question.

Category	p	h
PS	The cat sat on the mat.	The cat did not sit on the mat.
LE	The water is too hot.	The water is too cold.
MN	The new console is cheap.	The new console isn't cheap.
Q	Several are available.	All are available.

Table 1: Examples of NLI items from each AX category. MN and Q items have been trimmed and paraphrased to fit in one line, but still fall into their respective categories.

Our goal is to use LLMs to generate new items for AX, such that the psychometric properties of both the items and the test as a whole are improved. Formally, given an LLM M and a prompt p that contains one or more items that have a psychometric property θ, we seek to sample new items i from M that lead to an improvement in θ:[2]

$$i \sim M(p) \mid \theta_i > \theta_p \tag{1}$$

Where i and p are assumed to test for the same construct (e.g., NLI). Prior work has demonstrated that when LLMs are given existing items as prompts, they can generate new items that match the construct measured by those items (Liu et al., 2022; Lee et al., 2023). We build on this approach by designing prompts to instruct LLMs to generate new items for a particular construct, that *possess a desired psychometric property*. Figure 1 shows one of the prompts we developed. The model is instructed to generate only items that match the target property, and we use items from only one category at a time. We use item discrimination as the target property in our experiments. Discrimination refers to the ability of an item to separate high from low-ability test takers (Bandalos, 2018) and is computed using the item-to-total correlation (the correlation between the responses to a single item and total scores across all items). An item that is

[1] AX being the notation for the diagnostic on the GLUE leaderboard.

[2] Note that $\theta_i > \theta_p$ should be taken to mean that the psychometric properties of i are improved relative to p, and not necessarily that they are numerically greater.

```
I need to generate new NLI
items for a given trait.
Here are some examples:
###
Trait: High Discrimination
Items (3):
[ITEMS]
###
Trait: Low Discrimination
Items (3):
[ITEMS]
###
Trait: High Discrimination
New Items (5):
```

Figure 1: Prompt structure using the "simple" prompt format. Additional newlines have been added to keep text within margins.

highly discriminating will predict total scores and thus should be maximized. Our use of discrimination was based on preliminary analysis of the data from Laverghetta Jr. et al. (2021), which indicated that at least one item in every category had negative discrimination. In general, items with negative discrimination are regarded as problematic and possibly erroneous, and should not be included in cognitive assessments (Bandalos, 2018), which makes improving the discrimination of the AX items a natural optimization target. We use existing human written items as examples of the desired property in the prompt, selecting the top k items with the highest discrimination as "high discrimination" and the bottom k items with the lowest discrimination as "low discrimination".[3] We set $k = 3$ in our experiments, as we found larger values caused the difference in discrimination to become negligible. By providing examples of both good and bad items, we hope to teach the model general characteristics of high-quality items.[4]

We use GPT-3 (Brown et al., 2020) as our item generator, given its strong performance across many NLP tasks, the presence of an easy-to-use and inexpensive API, and the success prior work has had in using GPT-3 to generate non-cognitive (Lee et al., 2023) and NLI (Liu et al., 2022) items. We set temperature to 1 for all experiments, to encourage diversity in the generated items, and use a maximum token limit of 300. We explore the effect of varying other key hyperparameters:

- **Top P:** This parameter is based on nucleus sampling (Holtzman et al., 2019) and determines what fraction of log probabilities to consider when sampling, with larger values allowing more unlikely completions to be sampled. Prior work in LLM-based AIG has differed on this setting; some have used a value above 0.5 (Lee et al., 2023) and others a value at or below 0.5 (Liu et al., 2022). We therefore choose to experiment with both 0.5 and 1, as we theorized setting a higher value could lead to more diverse generations, but also increase the risk the items would lack construct validity.

- **Prompt Type:** We use a "simple" prompt following the structure shown in Figure 1. However, because the AX categories are highly specific, we reasoned that providing additional context about the categories may improve generation accuracy. We thus also experiment with "elaborated" prompts, which include additional information about each category, taken from the appendix on AX.[5]

We left all other hyperparamters at their defaults. We use the `text-davinci-003` endpoint,[6] and queried the API in December 2022. We generate 400 items, 100 for each category, and 25 for each hyperparameter combination (prompt type and top p). We remove any duplicate items, items where the model did not generate a valid label, and items that match verbatim an item from AX.

Following best practices in scale development (Worthington and Whittaker, 2006) we conduct a content review on the generated items. Four Ph.D. students with prior publications in NLP, NLI, or psychometric AI were asked to rate the quality of the GPT-3 items. We ask our annotators to rate the relevance of the items for measuring the category, the clarity of the items (in terms of whether they have spelling or grammatical errors), whether the items have potentially harmful content, and their

[3] Properties are calculated using SPSS version 28. We use only the categories from Table 1.

[4] Note that our approach has strong conceptual similarities to prior work in few-shot item selection for in-context learning (e.g. Walsh et al., 2022), in that the psychometric properties of the items are essentially used to select which shots to use.

[5] https://gluebenchmark.com/diagnostics

[6] Prompts and generated items for reproducing our results are available on Github: https://github.com/Advancing-Machine-Human-Reasoning-Lab/gpt3-item-generation/tree/main

certainty in their annotations. Before beginning the study, we gave annotators detailed instructions they were asked to review in advance, including information about the AX categories, how to answer each of the ratings, and example ratings. We instructed annotators to rate items as "Completely irrelevant" if either the label was incorrect or the item did not match the target category. We followed standard practices in NLI research for determining what the correct label should be (Bowman et al., 2015), which all our annotators were informed of. In particular, annotators always assumed p and h referred to the same event or situation (Bowman et al., 2015). For determining category membership, we follow the definitions of each AX category provided by Wang et al. (2018), and developed a simple code book for determining this. The majority of the annotations were done synchronously in a four-hour annotation session. Per recommended practices for content analysis, each item was rated by every annotator (Putka et al., 2008). Annotators were encouraged to discuss items with each other and come to an agreement on what ratings should be used. Further details on the content review, including an example of the annotation interface, can be found in Appendix A.

For a generated item to pass the content review, we determined that all annotators must rate the item as very clear, either relevant or very relevant, that the item contained no harmful content, and that annotators were either sure or very sure of their predictions. Of the 400 items, 92 met these criteria across all categories, with at least 15 in every category passing. We sampled 15 at random from each category, balanced for the label, to obtain the GPT-3 generated items. In total, 60 items were sampled.

4 Experiments

We determined in Section 3 that GPT-3 can generate AX items that possess at least face validity evidence. But are these items really more valid and reliable measures of basic linguistic reasoning, given that we designed our prompts to induce this? To study this, we recruited human participants on Amazon Mechanical Turk[7] to complete both the GPT-3 items and the original human-written items.

102 participants residing in the United States, who had completed at least 50 HITs (human intelligence tasks) with an acceptance rate of at least 90%, were recruited to take part in the study. We use the attention check items and quality control protocol from Laverghetta Jr. et al. (2021) to validate that our workers participated in good faith. Workers first completed an onboarding HIT where they were given five attention check items, whose format was identical to the AX items but by design, they were much easier to solve. This was meant to familiarize workers with the task and ensure they would likely give good response data. Workers who passed the onboarding then completed two more HITs, each containing half the GPT-3 items, and then two final HITs, each containing half the human-written items, and each of these HITs contained six attention checks spread evenly throughout the survey. Each worker's submission was evaluated on every survey, and we followed the protocols developed by Laverghetta Jr. et al. (2021) to determine whether work should be accepted or rejected. Briefly, workers needed to get at least 60% accuracy on the survey, or at least 66% on the attention checks, and provide a *justification* for each response to show that they were truly paying attention. Further details on the protocol and payment structure for the human studies are included in Appendix B.

We ultimately gathered data from 18 participants and base the following analysis on this sample. Broadly, our goal is to compare the psychometric properties of the GPT-3 written items to the human-written items, focusing specifically on item difficulty, item discrimination, reliability (assessed using internal consistency), and convergent and discriminant validity. These are all important properties to analyze when establishing the validity and reliability of a new assessment (Bandalos, 2018), and when assessed using a measurement framework known as classical test theory (CTT), can be computed using small sample sizes. CTT essentially posits that an individual's true proficiency on a cognitive task (their *true score*) can be decomposed into an observed (actual) score they obtain and an error term that represents the measurement error (Rust and Golombok, 2014). Note that this error is assumed to be random, and not systematic. Methods from CTT for assessing both validity and reliability are hence based on analysis of observed scores, and correlations between observed scores, where the observed scores are simply accuracy on the task:

$$observed\ score = \frac{correct\ answers}{all\ answers} \quad (2)$$

[7] https://www.mturk.com

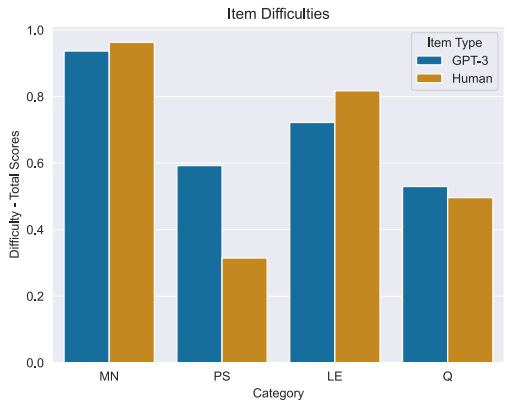

Figure 2: Mean item difficulties for each category, measured using total scores. Lower values indicate lower total scores, and hence more difficult items.

Although more sophisticated measurement theories have been developed (Embretson and Reise, 2013), they typically rely on latent variable modeling and require much larger sample sizes. Furthermore, in practice, establishing validity and reliability under CTT is often a first step in validating new assessments (Bandalos, 2018), which we believe justifies our focus on CTT in the present study.

4.1 Analysis of Item Properties

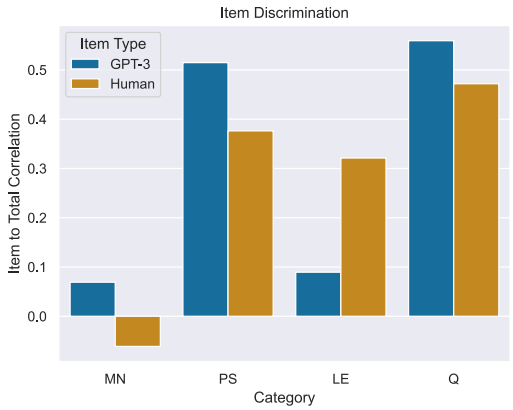

Figure 3: Mean item-to-total correlations for each category. Higher values indicate items are more predictive of a participant's total score, and hence are more discriminating.

We begin by comparing mean item difficulties (Figure 2) and mean item discriminations (Figure 3) for both human and GPT-3 written items. Difficulty is based on the participants' observed scores, and is equivalent to accuracy. Classical psychometrics dictates that items should have difficulties at approximately the midpoint between chance and

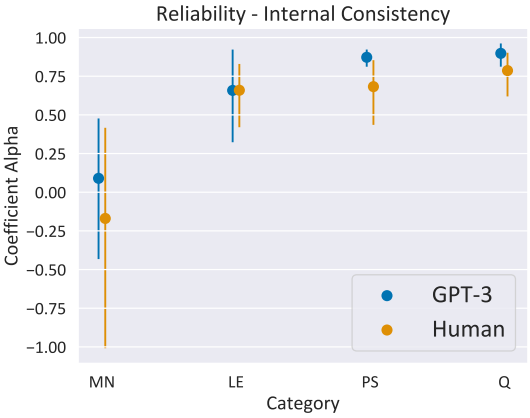

Figure 4: Coefficient α for item responses in each category, comparing human-written to GPT-3 written items. Errors bars are 95% confidence intervals computed using Feldt's method (Feldt et al., 1987). Higher values indicate better reliability and stronger validity evidence.

perfect scores (Lord, 1952), which in our case is roughly 70%. We again use item-to-total correlation to measure discrimination, and recall that item discrimination should be positive, with high values indicating better discrimination. We find that GPT-3 items are consistently closer to the optimal difficulty level than human-written items. GPT-3 items are also more discriminating than human-written ones, though a notable exception is for LE, where the GPT-3 items are noticeably less discriminating. As LE tests for all forms of lexical entailment, and is a much more broadly scoped construct than the others, lower discrimination is expected (Clark and Watson, 1995), though this does not fully explain the rather sizeable drop.

4.2 Internal Consistency Reliability

Items on cognitive assessments should exhibit strong reliability, meaning that participants with similar ability levels should also respond in a similar fashion. A widely used measure of reliability is coefficient α (Tavakol and Dennick, 2011), defined as:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_{y_i}^2}{\sigma_x^2}\right) \quad (3)$$

Where k is the total number of items, σ_x^2 is the variance of total scores across all items, and $\sigma_{y_i}^2$ is the variance of total scores for item i. α ranges from $-\infty$ to 1, and will be negative when there is greater within-subject variability than between-subject variability. Reliability should thus be maximized. We compute α for both GPT-3 and human

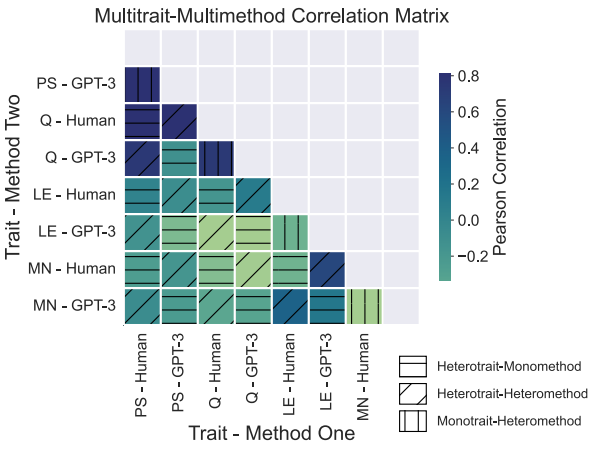

Figure 5: Results from the MTMM matrix, computed using Pearson correlations with total scores. Bluer colors indicate stronger correlation.

written items, doing so separately for each category, using the Pingouin Python library (Vallat, 2018). Reliabilites with 95% confidence intervals are shown in Figure 4. Across all categories, GPT-3 produces items with similar or better reliabilities compared to human-written items. MN is a special case, as α for this category dips into the negative range, indicating poor validity evidence, though even in this case the GPT-3 items show much better reliability overall. Thus, the GPT-3 items appear to elicit more consistent responses among human participants.

4.3 Convergent and Discriminant Validity Evidence

The multi-trait multi-method (MTMM) matrix is a classic technique for evaluating the construct validity of measures and is often used when evaluating new instruments (Campbell and Fiske, 1959). The MTMM matrix shows the correlations between different cognitive constructs (the traits) when they are measured using different measurement techniques (the methods). In this framework, validity is defined in terms of the strength of the correlation between different trait / method combinations. In general, different methods should be strongly correlated when measuring the same trait (monotrait-heteromethod), and different traits measured using the same method should be weakly correlated (heterotrait-monomethod), per the definitions of convergent and discriminant validity (Campbell and Fiske, 1959).

We use this approach to evaluate the convergent and discriminant validity of the GPT-3 items. We treat the AX category as the trait, and the method used to generate items (human written or generated by GPT-3) as the method and compute Pearson correlations between all possible combinations of trait and method, using the participant's total scores. Additionally, we check for significance using Bonferroni corrected p-values of 0.002.[8] Results are shown in Figure 5. Significant monotrait-heteromethod correlations were found for PS (ρ = 0.75, p $<<$ 0.001) but not for Q (ρ = 0.72, p $<$ 0.01), MN (ρ = 0.06, p $<$ 0.5) or LE (ρ = 0.20, p $<$ 0.5) All heterotrait-monomethod correlations were insignificant (p $>$ 0.1), except for between PS and Q. For human-written items, the correlation was found to be significant (ρ = 0.81, p $<<$ 0.001), but not for GPT-3 written items (ρ = 0.16, p $<$ 0.5). Collectively, these results indicate strong evidence for the discriminant validity of the GPT-3 items, given the lack of significant heterotrait-monomethod correlations. Evidence for convergent validity is strong for PS, and to a lesser extent Q,[9] but not for either MN or LE. Thus, the validity evidence for GPT-3 written items is just as strong, if not stronger, than for human-written items.

4.4 Analysis of Local Item Dependency

Recall that CTT assumes that measurement errors are due purely to random chance, and systematic error is not easily accounted for. One way this can be violated is from a phenomenon called local item dependence (LID). LID occurs between pairs of items, often whenever information needed to solve the items is interrelated. For example, LID is often a concern on reading comprehension assessments, because items that refer to the same text can inadvertently introduce local dependency on the common stimulus (Attali et al., 2022). Importantly, LID indicates that errors on items are interrelated in a way other than proficiency on the construct, and hence imply systematic error in the measurement.

As Attali et al. (2022) notes, LID is an even greater concern in the context of AIG, as GPT-3 may have generated items in a programmatic and somewhat redundant fashion. Perhaps as an artifact of how AX was constructed, we also found many human-written items had highly similar linguistic structures, which we reasoned could cause GPT-

[8] Rounded to three decimal places.

[9] The monotrait-hetheromethod correlations for Q were strong, even though they did not meet the Bonferrioni-corrected significance level.

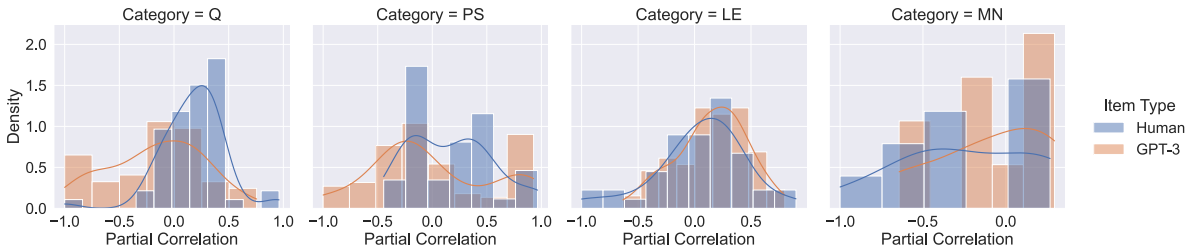

Figure 6: Density plots (computed using kernel density estimation) of partial Pearson correlations computed for each category, controlling for the participants' total scores per category. Item pairs where one or both items have 0 variance are excluded. Partial correlations greater than 0.3 indicate LID, and distributions which peak closer to 0 have fewer item pairs with LID.

3 to generate items based on a common stimulus, which might inadvertently introduce LID. We thus follow Attali et al.'s protocol and, for each category and for both the human-written and GPT-3 written items, we compute the partial correlations between all pairs of items in each category, controlling for total scores. Following prior work (Christensen et al., 2017; Attali et al., 2022), we use a threshold of 0.3 correlation or higher as indicating LID, and we plot the density distributions of the partial correlations in each category. Results are shown in Figure 6. We find that, even with the human-written items, LID appears to be present in all categories except for MN, though even in this case we observe strong anti-correlations. It does not appear, however, that the GPT-3 items have made LID significantly worse. Distributions are often similar between the item types, and in some cases, GPT-3 distributions appear closer to zero, indicating fewer pairs with LID. We thus surmise that LID is no greater a concern for GPT-3 written items than it was for human-written items.

4.5 Scaling Up to GPT-4

OpenAI's most recent LLM, GPT-4,[10] was released after the completion of our testing of the GPT-3 items. Given the large gains in performance reported for GPT-4 across myriad tasks, we chose to perform preliminary analysis on the quality of items generated by GPT-4, this time running only the content review.[11] We use the same content experts and follow an identical protocol for the review. We chose not to generate items for MN, due to the very poor validity evidence for items in this category. Hyperparameters and prompts remain the same, and we use the `gpt-4` endpoint in the API. To keep results as comparable as possible across models, we chose not to use the system context or other chat features provided for GPT-4, and instead administer the prompts in a single shot. We generate 18 items per category, totaling 54 across the three categories tested. After running deduplication and dropping items with invalid labels, we administer the remaining items to our content experts. We were specifically interested in whether our experts would report the GPT-4 items as being any more relevant for measuring the target construct as compared to GPT-3. We graph the annotator distributions for PS in Figure 7, and show results for LE and Q in Appendix C. Surprisingly, we find results from GPT-4 to be mixed. Although GPT-4 generates a larger fraction of items labeled as either "Relevant" or "Very relevant" for Q, it generates fewer such items for LE and PS. As GPT-4 is designed to function more like a chatbot than GPT-3, it is possible our prompts need to be restructured to make better use of the model's capabilities, but more experiments are needed to explore this.

5 Discussion and Conclusion

Collectively, our results demonstrate that LLMs can generate items with superior validity evidence, even for constructs that have undergone limited psychometric analysis. GPT-3 items were found to have better discrimination and reliability, while maintaining strong convergent, discriminant, and content validity. LID, while confirmed to be present in both item types, appeared no worse and perhaps slightly better in GPT-3 items. These positive results, while clearly present for PS and Q, were less clear for MN and LE, and validity evidence as a whole appeared strongest for the categories testing the most narrowly scoped constructs.

[10] https://openai.com/research/gpt-4

[11] Due to time constraints, we could not run a more detailed analysis on the GPT-4 written items, and leave this to future work.

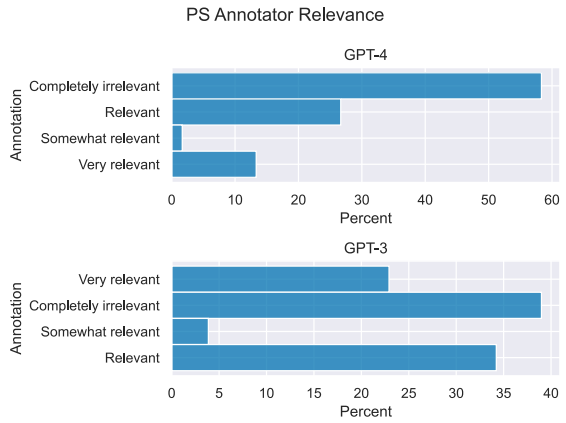

Figure 7: Distribution of annotator relevance scores (checking that the item both has a correct label and matches the category) for both GPT-3 and GPT-4 items, on items from the PS category. A lower percentage of items marked as "Completely irrelevant" indicates stronger evidence of the content validity of items generated using that model.

Though promising, our results come with limitations that should be addressed in future work. The small sample size we collected makes it difficult to assess the generalizability of our findings. This also prevented us from running any analysis of internal structure or differential item functioning (DIF) using methods from factor analysis or item response theory, as these models require large sample sizes (Min and Aryadoust, 2021). As items generated by GPT-3 should contain no DIF and have similar factor structures as items written by humans, these are important analyses to explore in future work. We also did not examine the *diversity* of the generated items, in other words, how thoroughly the model explored the construct space. It is a well-known problem in psychometrics that having too many similarly worded items can inflate the reliability and reduce the validity of a measure (Clark and Watson, 1995), and our results may have been susceptible to this. A related problem is ensuring that the distribution of labels in the generated items remains balanced, and while we took steps to account for this, we did find that the distribution of GPT-3 items was somewhat unbalanced. For example, there were far fewer neutral items than either entailment or contradiction. Improving the prompt design to account for diversity and other psychometric properties simultaneously is a fruitful direction for future work. Our experiment with GPT-4, while disappointing, was also quite limited and should be expanded upon. We deliberately kept the prompt design as similar as possible between the two models, to avoid possible confounds. Making effective use of the system query and changing the structure of the prompts to suit a conversational style could lead to much better results, however. Finally, although we believe NLI is a good task to use for initial experimentation, we also acknowledge that it is significantly different from the tasks of interest in education (e.g., question answering), and future work should explore our approach on tasks with stronger educational applications.

LLMs have the potential to greatly ease the burden of scale development, and transform educational and psychological measurement. Our results contribute to the growing field of LLM-based automated item generation, and demonstrate the potential these methods have for generating valid and reliable items at a scale that would have previously been impossible. Further research, combining our approach with more advanced prompting strategies, or zero-shot parameter estimation, could conceivably lead to a system that generates high-quality items in a fully autonomous fashion, which would transform the practice of writing and validating test items.

Limitations

We emphasize that our research is exploratory and the generated items we produced should not be used for making critical evaluations of cognitive skillsets in either humans or LLMs. As discussed in Section 5, our small sample size makes it difficult to draw broad conclusions about the generalizability of our findings, and practical considerations regarding the annotation study limited our ability to thoroughly explore the prompt space. While we chose GPT-3 due to its ease of use and the fact that most psychometricians would likely be aware of it, we also acknowledge that OpenAI has released few details on how this model is trained or updated, which hampers the reproducibility of our results. We also acknowledge that more recent OpenAI LLMs, including ChatGPT and GPT-4, have been released since this work is completed, and that our preliminary experiments using GPT-4 do not give us a full understanding of the capabilities of this model. However, given that we were still able to perform detailed experiments using the GPT-3 items, and these items proved to have superior validity evidence across multiple trials, we do not believe the existence of more recent LLMs negates

our results. Finally, it is also well known that LLMs can produce biased, toxic, or other forms of harmful text content (Liang et al., 2021). While we took steps to account for this in our content review, future work must keep this possibility in mind and carefully analyze generated items for potentially harmful content. A related problem is the risk of GPT-3 items propagating disadvantages against historically marginalized groups. For example, the items may have relied on cultural context or other information that would give an unfair advantage to certain populations. Given that we lacked a sufficient sample size and did not collect personally identifiable information from participants, we could not run DIF analysis to check for this, and cannot state definitively that DIF is not present.

Acknowledgements

We would like to thank Logan Fields, Animesh Nighojkar, and Zaid Marji for assisting us with the content review. Part of this research was sponsored by the DEVCOM Analysis Center and was accomplished under Cooperative Agreement Number W911NF-22-2-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Ahmed M Asfahani. 2022. The impact of artificial intelligence on industrial-organizational psychology: A systematic review. *The Journal of Behavioral Science*, 17(3):125–139.

Yigal Attali, Andrew Runge, Geoffrey T. LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A. von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5.

Deborah L Bandalos. 2018. *Measurement theory and applications for the social sciences*. Guilford Publications.

Isaac I Bejar, René R Lawless, Mary E Morley, Michael E Wagner, Randy E Bennett, and Javier Revuelta. 2002. A feasibility study of on-the-fly item generation in adaptive testing. *ETS Research Report Series*, 2002(2):i–44.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Donald T Campbell and Donald W Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2):81.

Sophia Chan, Swapna Somasundaran, Debanjan Ghosh, and Mengxuan Zhao. 2022. Agree: A system for generating automated grammar reading exercises. *arXiv preprint arXiv:2210.16302*.

Xieling Chen, Haoran Xie, Di Zou, and Gwo-Jen Hwang. 2020. Application and theory gaps during the rise of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 1:100002.

Song Cheng, Qi Liu, Enhong Chen, Zai Huang, Zhenya Huang, Yiying Chen, Haiping Ma, and Guoping Hu. 2019. Dirt: Deep learning enhanced item response theory for cognitive diagnosis. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2397–2400.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Karl Bang Christensen, Guido Makransky, and Mike Horton. 2017. Critical values for yen's q 3: Identification of local dependence in the rasch model using residual correlations. *Applied psychological measurement*, 41(3):178–194.

Lee Anna Clark and David Watson. 1995. Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3):309.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First*

PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark. Association for Computational Linguistics.

Susan E Embretson and Steven P Reise. 2013. *Item response theory*. Psychology Press.

Leonard S Feldt, David J Woodruff, and Fathi A Salih. 1987. Statistical inference for coefficient alpha. *Applied psychological measurement*, 11(1):93–103.

Friedrich M Götz, Rakoen Maertens, Sahil Loomba, and Sander van der Linden. 2023. Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*.

Ivan Hernandez and Weiwen Nie. 2022. The ai-ip: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Antonio Laverghetta Jr., Animesh Nighojkar, Jamshidbek Mirzakhalov, and John Licato. 2021. Can transformer language models predict psychometric properties? In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 12–25, Online. Association for Computational Linguistics.

Philseok Lee, Shea Fyffe, Mina Son, Zihao Jia, and Ziyu Yao. 2023. A paradigm shift from "human writing" to "machine generation" in personality test development: an application of state-of-the-art natural language processing. *Journal of Business and Psychology*, 38:163–190.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Frederic M Lord. 1952. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika*, 17(2):181–194.

Rakoen Maertens, Friedrich Götz, Claudia R Schneider, Jon Roozenbeek, John R Kerr, Stefan Stieger, William Patrick McClanahan III, Karly Drabot, and Sander van der Linden. 2021. The misinformation susceptibility test (mist): A psychometrically validated measure of news veracity discernment.

Shangchao Min and Vahid Aryadoust. 2021. A systematic review of item response theory in language assessment: Implications for the dimensionality of language ability. *Studies in Educational Evaluation*, 68:100963.

Elham Mousavinasab, Nahid Zarifsanaiey, Sharareh R. Niakan Kalhori, Mahnaz Rakhshan, Leila Keikha, and Marjan Ghazi Saeedi. 2021. Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, 29(1):142–163.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Animesh Nighojkar and John Licato. 2021. Improving paraphrase detection with the adversarial paraphrasing task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7106–7116, Online. Association for Computational Linguistics.

Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, and Matthias Samwald. 2022. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1):6793.

Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. 2015. Deep knowledge tracing. *Advances in neural information processing systems*, 28.

Septian Eko Prasetyo, Teguh Bharata Adji, and Indriana Hidayah. 2020. Automated item generation: Model and development technique. pages 64–69. IEEE.

Dan J Putka, Huy Le, Rodney A McCloy, and Tirso Diaz. 2008. Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, 93(5):959.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Manav Rathod, Tony Tu, and Katherine Stasaski. 2022. Educational multi-question generation for reading comprehension. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 216–223, Seattle, Washington. Association for Computational Linguistics.

Cristobal Romero and Sebastian Ventura. 2020. Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1355.

John Rust and Susan Golombok. 2014. *Modern psychometrics: The science of psychological assessment*. Routledge.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951.

Stephen Stark, Oleksandr S Chernyshenko, and Nigel Guenole. 2011. Can subject matter experts' ratings of statement extremity be used to streamline the development of unidimensional pairwise preference scales? *Organizational Research Methods*, 14(2):256–278.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Mohsen Tavakol and Reg Dennick. 2011. Making sense of cronbach's alpha. *International journal of medical education*, 2:53.

Mikke Tavast, Anton Kunnari, and Perttu Hämäläinen. 2022. Language models can generate human-like self-reports of emotion. In *27th International Conference on Intelligent User Interfaces*, pages 69–72.

Raphael Vallat. 2018. Pingouin: statistics in python. *J. Open Source Softw.*, 3(31):1026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Matthias von Davier. 2018. Automated item generation with recurrent neural networks. *Psychometrika*, 83:847–857.

Reece Walsh, Mohamed H Abdelpakey, Mohamed S Shehata, and Mostafa M Mohamed. 2022. Automated human cell classification in sparse datasets using few-shot learning. *Scientific Reports*, 12(1):2924.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Roger L Worthington and Tiffany A Whittaker. 2006. Scale development research: A content analysis and recommendations for best practices. *The counseling psychologist*, 34(6):806–838.

Bowei Zou, Pengfei Li, Liangming Pan, and Ai Ti Aw. 2022. Automatic true/false question generation for educational purpose. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 61–70, Seattle, Washington. Association for Computational Linguistics.

Jiyun Zu, Ikkyu Choi, and Jiangang Hao. 2023. Automated distractor generation for fill-in-the-blank items using a prompt-based learning approach. *Psychological Testing and Assessment Modeling*, 65(2):55–75.

A Details on Content Review

Content review ratings were collected via Qualtrics.[12] We developed five items to ask our experts:

[12] https://www.qualtrics.com

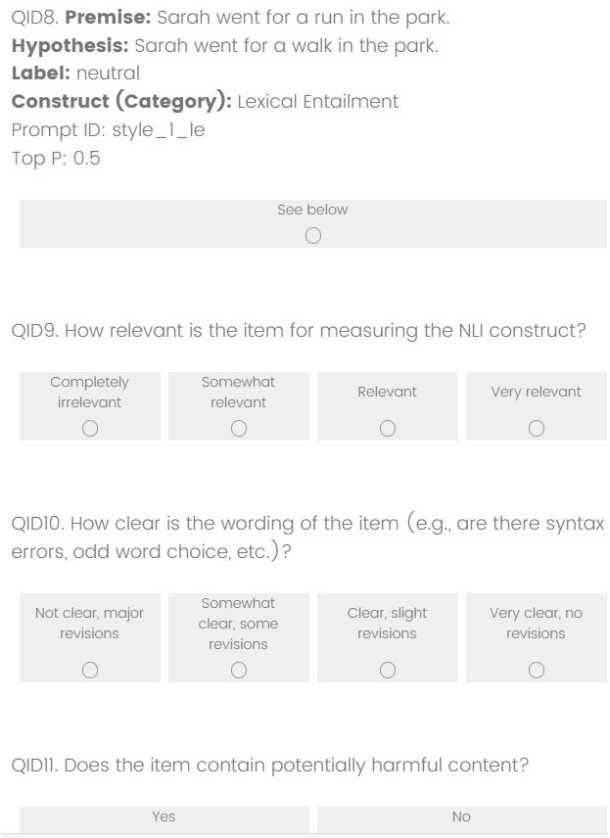

Figure 8: The annotation interface for the content review.

1. **Item Relevance:** This question concerned the usefulness of the item for measuring the construct. Experts could rate items as "Completely irrelevant", "Somewhat relevant", "Relevant", or "Very relevant". At a basic level, items needed have both a correct label and test for the target category. If *either* of these were false, experts were instructed to rate the item as "Completely irrelevant". Experts were instructed to rate items as "Somewhat relevant" if the prior checks passed, but knowledge of the category was not critical to solving the item. An example of this would be an item from MN where the negated clause does not change at all from p to h. If knowledge of the category was critical, and all prior checks passed, experts were instructed to rate the item as "Relevant". "Very relevant" was reserved for items that experts judged as being highly discriminating, which we included based on prior work demonstrating experts can effectively evaluate latent properties of items (Stark et al., 2011). We left the exact judgment of what constituted a highly discriminating item up to the discretion of the experts, and we encouraged them to discuss this and reach an agreement for each item deemed "Very relevant".

2. **Item Clarity:** This question concerned how clear the wording of the item is, and whether it contains spelling or grammatical errors. Experts could rate items as "Not clear, major revisions", "Somewhat clear, some revisions", "Clear, slight revisions", and "Very clear, no revisions". "Not clear, major revisions" was reserved for cases where items contained any spelling or grammatical errors. This also included cases with unterminated punctuation (e.g, an opening '(' that was not closed). Both "Somewhat clear, some revisions" and "Clear, slight revisions" were reserved for cases where the prose of the item was unorthodox (e.g, GPT-3 generated an odd word choice or an unusual phrase). Experts were instructed to rate "Very clear, no revisions" if items were both grammatically correct and contained no unusual wording that made the item needlessly difficult to understand.

3. **Potentially Harmful Content:** This was included to ensure that GPT-3 did not generate offensive or otherwise harmful content in the items, though we did not expect this to be an issue in general as AX items were written in a fairly neutral tone and avoided covering controversial social issues or explicitly targeting identified subgroups. Experts were instructed to check if the items contained any content related to race, ethnicity, religion, or other identifiable characteristics that might be considered offensive to members of those groups. Importantly, AX does contain items related to U.S. politics circa 2018 that we reasoned might lead to toxic generations regarding political ideology. We made experts aware of this but instructed them to *only* rate such items as harmful if the content explicitly attacked a political ideology or its adherents. There were only two options for this item, "yes" or "no".

4. **Annotator Certainty:** Finally, using a four-point Likert scale, we asked annotators to rate how sure they were of their ratings.

Figure 8 shows the annotation interface. Experts were given the full item content and the label generated by GPT-3, as well as additional data about the hyperparameters used which they did not need to refer to. They were free to move back and forth within the survey and revise their responses later if they wished. Most annotations were completed in a synchronous session, and all annotators began their work in this session to ensure the task instructions were clear and to train them on how to rate each item. Importantly, we did not ask raters to edit any item content to improve its quality, as we were interested in the quality of GPT-3 written items without human intervention.

For determining category membership, we developed a codebook based on the presence of certain keywords in the item content, and either p or h needed to contain at least one of these keywords to pass content validity. For example, for Q, either p or h needed to contain either a universal (all, none) or existential (some, many, most, etc.) quantifier in natural language to pass. We developed an initial list of keywords based on both the appendix covering AX in Wang et al. (2018), and by manually inspecting the items in each category to locate additional keywords. During the content review, experts could also suggest additional keywords, and if all annotators agreed, these new keywords were added to the codebook. Table 2 show all the keywords used across categories. LE was the only category that did not follow this protocol for determining category membership. As LE tests for all forms of entailment at the word level, there is no predetermined list of keywords that can be used to determine LE membership. Therefore, for LE, we used the rule that p and h must differ by only one word, with the only exception being if other words needed to be changed to keep the sentences grammatically correct.

B Details on Human Study

We follow many of the same protocols from Laverghetta Jr. et al. (2021) for conducting our human study. In particular, they employed attention check NLI items taken from the ChaosNLI dataset (Nie et al., 2020), which collected 100 human ratings to a subset of SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018) items. Only items which at least 90% of the workers agreed on the correct label were used, and hence they are presumably quite easy to answer correctly. In addition, Laverghetta Jr. et al. (2021) also asked workers to justify their response to each item, which was used as an additional check to ensure workers were paying attention during the task. We follow their protocol and check that workers do not copy text from the item as their justification, that the justification is not used multiple times, that it is clearly related to the item content,[13] and that the justification is not a nonsensical word or phrase (e.g, "good" or "nice question"). Collectively, the following quality control procedure was used for each survey:

1. Submissions with duplicate IP addresses or worker IDs were dropped.

2. Submissions with less than 40% accuracy, or less than 60% with less than 66% on attention checks, were dropped.

3. Submissions whose justifications did not meet the above criteria were also dropped.

All other submissions were accepted, and at each stage passing workers were given qualifications to proceed to the next survey. If however, workers

[13] In some instances, workers appeared to copy text from external websites that was completely unrelated to the question.

Category	Keywords
MN	un−, non− ir−, dis−, im−, il−, in−, −n't, not, never, no
PS	un−, non− ir−, dis−, im−, il−, in−, −n't, not, no, and, or, if
Q	all, no, some, many, most, none, every, several, each, one other, only, nearly all, the , part of

Table 2: Keywords used to determine category membership. Leading and trailing "−" indicate suffixes and prefixes, respectively.

failed a given stage, they were not allowed to proceed. In total, we administered five separate HITs and used Qualtrics to gather all responses. Workers were paid $8.00 for each HIT, except for the initial onboarding HIT, where they were paid $0.10,[14] and had one hour to complete each HIT. Workers were told they would be compensated for each survey completed successfully, to encourage consistently high-quality work. Workers gave informed consent to participate prior to beginning each HIT, and could withdraw at any time. Workers could appeal any rejections made, however, we also clearly stated submissions would be checked for quality control purposes, and may be dropped if evidence of bad-faith responses was found. All work was done anonymously; workers were not asked to provide us with any personally identifiable information at any stage.

Finally, we also considered extending Laverghetta Jr. et al.'s protocol to check for AI-generated text for the explanations, in case workers attempted to use ChatGPT or another LLM during the survey. We examined several detectors for AI-written text, including one developed by OpenAI.[15] However, we found that currently available models require too much text to be helpful for our study. Participants were asked to only briefly explain their thought process with at most one sentence, which was far too short for current detectors to make a classification. Therefore, we did not include any check for AI-generated text, but we strongly encourage future work to consider this and investigate other possible safeguards against workers cheating on the task using LLMs.

C Additional Results from GPT-4

Figures 9 and 10 compare the annotator relevance scores between GPT-3 and GPT-4 items, for LE and Q.

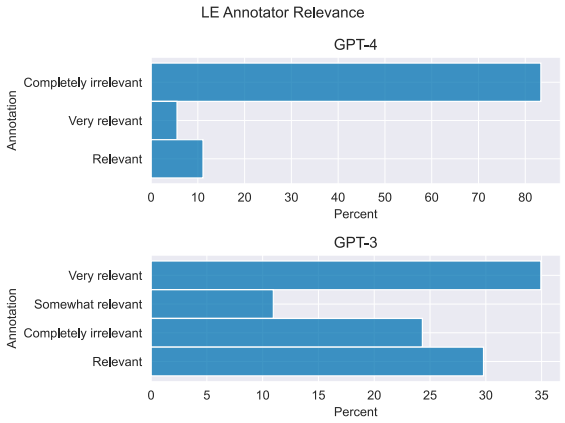

Figure 9: Distribution of annotator relevance scores for LE.

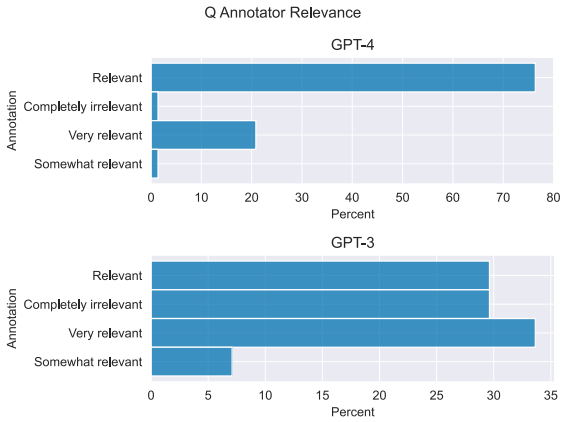

Figure 10: Distribution of annotator relevance scores for Q.

[14] This HIT contained only 5 items and was meant to be finished quickly.

[15] https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text