A. Details on prompt engineering

A.1. Item generation

Figure 8 lists the few-shot item generation prompt. Text in brackets indicates template strings that are filled by LangChain. We include the instructions given to annotators during the content review, as we reason that these criteria would provide helpful context for the characteristics of good items. The prompt includes human-written CPS exemplars in addition to the LLM generations selected by the item evaluator. We find including these exemplars helps reduce subtle content errors, like priming effects. Unlike the LLM written exemplars, which update at every iteration, the human items always remain fixed and are selected by the lead authors from the items used in [1].

B. Details on word list generation and evaluation

Figure 9 lists the prompt we use.

C. Details on constraint satisfaction

Across all trials, we set both δ_o and δ_v to 0.02, as we found in preliminary trials that this setting was effective at selecting exemplars that improved originality without a drastic increase in similarity. It is worth noting that, in some cases, there may not be a η that satisfies the constraints. To account for this, we additionally include parameters γ_o and γ_v to decrement and increment δ_o and δ_v respectively. Because the range for originality is roughly double that of similarity (-2 to 2 vs. -1 to 1), we set γ_o to 0.002 and γ_v to 0.001, which again is based on preliminary trials. γ progressively relaxes the constraints of the optimization; for every iteration that fails to find an acceptable η, we apply the γ to δ update and try again. This continues until we reach a base case of minimum originality and maximum similarity, which, by definition, any η should satisfy (though in practice, this base case was never hit). This allows us to account for poor generations at any iteration, with the downside being that we sometimes see large increases in similarity while running constraint satisfaction. Though we find in practice this solution is effective, we believe that further improvements can be made. Neither δ nor γ were tuned as hyperparameters, which should be explored in future work. Likewise, there exist other possible solutions to address the shortcomings of greedy selection, like using a variable size η, or using more sophisticated prompting strategies designed to address diversity. However, because our goal behind item generation is not necessarily to find the most optimal pool of items, we leave these experiments to future work.

D. Full listing of `CPIG` runs

Table 4 lists all `CPIG` trials we performed. We attempt to balance all hyperparameters as evenly as possible across trials, although we ultimately ended up with somewhat uneven groups of trials (more greedy shot selection trials than either constraint satisfaction or random baseline). However, because each condition includes multiple runs, we do not believe this imbalance affects the significance of our results.

E. Example LLM item responses

We include two examples of responses to items generated by LLama-13b in Table 2. We find that the LLMs typically generate high-quality responses that obey our task instructions, and this remains consistent across iterations. Item responses also generally follow the patterns typically observed in humans, with the notable exception that they are much more elaborate, often including multiple actions to pursue as opposed to one idea. Given that all our LLMs are trained as chat-bots, it makes sense that they provide more elaboration to be as helpful as possible, though this remains an important discrepancy that future work should address. Another notable pattern is the tendency of LLMs to role play as one of the characters in the scenario when generating a solution, exemplified by the round one response. We do not observe this behavior in the data from [1], as participants are not instructed to do this, though we also did not explicitly instruct LLMs on how to provide their solution. While it is unclear what effect such behavior would have on the originality of LLMs' responses (if any), we believe it would be an interesting topic to explore in future work.

An example human-written CPS item is provided in Table 3. This is one of the items used to gather human responses for the scoring model from [1], and we also use it as one of our few-shot exemplars to seed item generation.

Table 2
Example responses for items generated by LLAMA-13B, from the first (top) and last (bottom) rounds of generation.

Creative Response
I would like to speak with Mark about the issues that Noah has been causing at the beach. I understand that Noah is a friend of mine, but his behavior is unacceptable and affecting the success of the business. I would suggest that we have a meeting with Noah and discuss the problems he has been causing, and see if there is a solution that can be reached. If Noah is unwilling to change his behavior, then perhaps it would be best for him to find a different job. I am willing to do whatever it takes to make sure that the beach is successful, even if it means having an uncomfortable conversation with my friend.
Elijah should have an open and honest conversation with Olivia about the situation. He should explain to her that he values their relationship and doesn't want to cause any problems, but he also doesn't want to condone Noah's behavior and risk getting in trouble with the theater's management. He should also let her know that he is torn between his loyalty to his friends and his commitment to doing the right thing. By being open and honest, Elijah can show Olivia that he is a trustworthy and responsible partner, and that he is willing to make difficult decisions when necessary. Additionally, he should also talk to Noah and let him know that his behavior is not acceptable and that it could have consequences. By addressing the situation directly, Elijah can show that he is a responsible and caring person, and that he is committed to doing the right thing, even if it is difficult.

Table 3
The "Becky" scenario used in the human studies from [1].

Human-Written CPS Item
Becky is a college student who works part-time at Mark's Pizzeria. Mark, the owner of the restaurant, has treated Becky very well. He gave her a job that she needs to help pay her rent when no other business would employ her because she was arrested for shoplifting three years ago. Mark also lets Becky work around her school schedule, and has asked if she wants to be a shift manager in the summers. Becky's roommate Jim also works at the pizzeria, but Jim has been causing a lot of problems at work. He always avoids doing his job, treats customers rudely, and makes a lot of mistakes with orders. Jim recently began stealing food from the pizzeria. Two days ago the pizzeria was short-staffed, so Jim and Becky were the only employees left at closing time. Jim made 10 extra pizzas and took them home to a party he was hosting without paying for them. Becky feels like she needs to do something about Jim's behavior. However, Becky is hesitant to tell Mark about Jim because Jim is a good friend to Becky. Becky also needs Jim to have a job so he can pay his portion of their rent. Becky does not know what to do.

```
SYSTEM:
You are an author tasked with producing scenarios for a short story. You
    will be given
a list of 5 words, consisting of 3 names, a place, and an action. Using
    ONLY these words,
think of a scenario that involves all the words. This scenario should
    involve a dilemma
that one of the named people from the list, the main character, needs to
    solve. At the
end of your scenario, write ‘‘I am finished with this scenario.’’ UNDER NO
    CIRCUMSTANCES
SHOULD YOU STATE WHAT THE MAIN CHARACTER SHOULD DO, HAS TO DO, IS GOING TO
    DO, OR WANTS TO
DO. LEAVE ALL POSSIBLE SOLUTIONS AMBIGUOUS. DO NOT ASK RHETORICAL QUESTIONS
     ABOUT WHAT THE
MAIN CHARACTER SHOULD DO. Here is a list of rules you should follow when
    writing the scenario:
[ANNOTATION INSTRUCTIONS]
###

Here are examples of a high-quality scenario that follows all of these
    rules:
[HUMAN WRITTEN ITEMS]

###

 Now write a new scenario that is similar to these, using the wordlist you
     will be provided.

 USER:
 Here are some more examples of high quality scenarios from other authors.
     Use these scenarios as guidance,
 but avoid drawing from them too heavily when developing your own:
 [LLM WRITTEN ITEMS]

 Word list:
[WORD LIST]

###

Scenario:
```

Figure 8: Item generation prompt. Text in red delineates the start of system and user context, respectively. Text in brackets are template strings filled by LangChain.

```
SYSTEM:
You are an author tasked with coming up with scenarios for a short story.
    Create a list of 5 words.
In the list, include 3 human names, a place, and an action. You don't need
    to use names of actual
people or places, but only use words which relate to the experiences of
    typical people; for example,
do not include a word like ``starship'' for the place since no one alive
    today has been on a starship.
Each entry in the list should consist of only a single word You should list
     words in exactly this order:
name, place, name, action, name.

 USER:
 Create 10 wordlists, make sure to use a different action each time.
     Separate wordlists by two newlines, do not number them.
```

Figure 9: The word list generation prompt.

```
SYSTEM:
You are a participant in an experiment. You will be presented with a
    problem scenario,
and must come up with a solution to the problem. Be creative in your
    response, but keep
it at no more than 4 sentences in length. Respond in a single paragraph.

 USER:
 Scenario:
[CREATIVE SCENARIO}
```

Figure 10: The no context item response generation prompt. Text in red delineates the start of system and user context, respectively. Text in brackets are template strings filled by LangChain.

```
SYSTEM:
You are a participant in an experiment. You are a [ETHNICITY] [GENDER] who
    works in [INDUSTRY].
Your job title is [TITLE]. You will be presented with a problem scenario,
    and must come up with
a solution to the problem. Be creative in your response, but keep it at no
    more than 4 sentences
in length. Respond in a single paragraph.

 USER:
 Scenario:
[CREATIVE SCENARIO}
```

Figure 11: The demographic item response generation prompt. Text in red delineates the start of system and user context, respectively. Text in brackets are template strings filled by LangChain. Demographic data is pulled from prior human responses to CPS items reported in [1].

```
SYSTEM:
You are [FIRST NAME] [LAST NAME], a participant in an experiment. You are a
     [OCCUPATION]
who works in [INDUSTRY]. [PSYCHOMETRIC DATA]. You will be presented with a
    problem scenario, and
must come up with a solution to the problem. Be creative in your response,
    but keep it
at no more than 4 sentences in length. Respond in a single paragraph.

 USER:
 Scenario:
[CREATIVE SCENARIO}
```

Figure 12: The psychometric item response generation prompt. Text in red delineates the start of system and user context, respectively. Text in brackets are template strings filled by LangChain. Psychometric data is pulled from prior human responses to CPS items reported in [1].

Table 4
Summary of hyperparameters for all CPIG trials. To the fullest extent possible, we balanced the hyperparameters such that each was run multiple times to ensure reproducibility. Some smaller LLMs failed to generate any items that passed content validity after multiple attempts and were excluded. The seed hyperparameter specifies the random seed used during inference.

Seed	Item gen model	Item response model	Shot selection method	Response prompt type
999	LLama-13b	LLama-13b	greedy	Demographic
333	LLama-13b	LLama-13b	greedy	Demographic
777	LLama-13b	LLama-13b	greedy	Demographic
333	Vicuna-13b	Vicuna-13b	greedy	Demographic
777	Vicuna-13b	Vicuna-13b	greedy	Demographic
999	Vicuna-13b	Vicuna-13b	greedy	Demographic
999	LLama-7b	LLama-7b	greedy	Demographic
777	LLama-7b	LLama-7b	greedy	Demographic
333	LLama-7b	LLama-7b	greedy	Demographic
777	LLama-70b	LLama-70b	greedy	Demographic
333	LLama-70b	LLama-70b	constraint satisfaction	Demographic
777	LLama-70b	LLama-70b	constraint satisfaction	Demographic
777	Vicuna-13b	Vicuna-13b	constraint satisfaction	Demographic
999	Vicuna-13b	Vicuna-13b	constraint satisfaction	Demographic
999	Vicuna-7b	Vicuna-7b	greedy	Demographic
777	Claude-3-haiku	LLama-7b	constraint satisfaction	Psychometric
333	Claude-3-haiku	LLama-7b	constraint satisfaction	Psychometric
999	Claude-3-haiku	LLama-7b	random	Psychometric
777	Claude-3-haiku	LLama-7b	random	Psychometric
333	Claude-3-haiku	LLama-7b	random	Psychometric
999	Vicuna-7b	Vicuna-7b	constraint satisfaction	Demographic
333	LLama-13b	LLama-13b	constraint satisfaction	Psychometric
777	LLama-13b	LLama-13b	constraint satisfaction	Psychometric
999	LLama-13b	LLama-13b	constraint satisfaction	Psychometric
333	Vicuna-13b	Vicuna-13b	constraint satisfaction	Psychometric
777	Vicuna-13b	Vicuna-13b	constraint satisfaction	Psychometric
999	Vicuna-13b	Vicuna-13b	constraint satisfaction	Psychometric
333	LLama-13b	LLama-13b	constraint satisfaction	No context
777	LLama-13b	LLama-13b	constraint satisfaction	No context
999	Vicuna-13b	Vicuna-13b	constraint satisfaction	No context
333	LLama-13b	LLama-13b	Random	Demographic
777	LLama-13b	LLama-13b	Random	Demographic
999	LLama-13b	LLama-13b	Random	Demographic
777	LLama-70b	LLama-70b	Zero shot	Demographic
333	Vicuna-13b	Vicuna-13b	Zero shot	Demographic
777	Vicuna-13b	Vicuna-13b	Zero shot	Demographic
999	Vicuna-13b	Vicuna-13b	Zero shot	Demographic
999	LLama-70b	LLama-70b	constraint satisfaction	Demographic
999	Claude-3-haiku	LLama-7b	Zero shot	Psychometric