

# Can LLMs evaluate items measuring collaborative problem-solving?

Ella Anghel<sup>1,\*</sup>, Yu Wang<sup>2</sup>, Madhumitha Gopalakrishnan<sup>2</sup>, Pranali Mansukhani<sup>2</sup> and Yoav Bergner<sup>2</sup>

<sup>1</sup>International Study Center, Lynch School of Education and Human Development, Boston College, 140 Commonwealth Ave, Chestnut Hill, MA 02467, USA

<sup>2</sup>Department of Administration, Leadership & Technology, New York University Steinhardt School of Culture, Education & Human Development, 82 Washington Square East, New York, NY 10003, USA

## Abstract

Collaborative problem-solving (CPS) is a vital skill for students to learn, but designing CPS assessments is challenging due to the construct's complexity. Advances in the capabilities of large language models (LLMs) have the potential to aid the design and evaluation of CPS items. In this study, we tested whether six LLMs agree with human judges on the quality of items measuring CPS. We found that GPT-4 was consistently the best-performing model with an overall accuracy of .77 ( $\kappa = .53$ ). GPT-4 did the best with zero-shot prompts, with other models only marginally benefiting from more complex prompts (few-shot, chain-of-thought). This work highlights challenges in using LLMs for assessment and proposes future research directions on the utility of LLMs for assessment design.

## Keywords

large language models, item evaluation, collaborative problem-solving, prompt engineering

## 1. Introduction

Collaborative problem-solving (CPS) is one of the most important 21st century skills according to employers [1] and has for some time attracted the interest of K-12 educators and policy-makers. High-quality assessment of CPS is a vital companion for curricula designed to develop this skill. However, the complexity of the construct makes it challenging to design items that properly target CPS and to evaluate the quality of candidate items. In recent years, the use of large language models (LLMs) and other AI-based methods were proposed for determining psychometric properties such as item difficulty [2]. These approaches are rarely applied to the evaluation of items' construct representation or to complex constructs like CPS. Therefore, it is unclear whether LLMs are suitable for such tasks. The current study aims to fill this gap by testing whether LLMs agree with humans on quality criteria for CPS items.

---

*EvalLAC'24: Workshop on Automatic Evaluation of Learning and Assessment Content, July 08, 2024, Recife, Brazil*

\*Corresponding author.

✉ anghel@bc.edu (E. Anghel); yw3060@nyu.edu (Y. Wang); mg7584@nyu.edu (M. Gopalakrishnan); pm3598@nyu.edu (P. Mansukhani); yoav.bergner@nyu.edu (Y. Bergner)

🆔 0000-0001-6332-7826 (E. Anghel); 0009-0005-9647-3076 (Y. Wang); 0009-0001-4328-5910 (M. Gopalakrishnan); 00009-0005-5801-4266 (P. Mansukhani); 0000-0001-7738-4290 (Y. Bergner)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

## 2. Literature review

### 2.1. Collaborative learning

It is now well established that collaboration and teamwork are essential for success in educational and work settings [1, 3]. The importance of collaborative problem solving (CPS) has led policy-makers to advocate for the development of high-quality CPS assessments [4]. These calls have been answered by several national and international assessment programs [5].

From a socio-cognitive perspective, CPS is also believed to improve learning of the underlying domain. However, simply working together on a task is not enough to facilitate learning [6]. Good CPS tasks should be challenging enough to justify the higher cognitive load of collaborating [7], focus on conceptual rather than procedural material [8], and involve positive interdependence among participants [9, 10].

While the importance of CPS in and of itself and as a contributor to other learning is widely supported, it remains a difficult construct to assess [11, 12]. Some challenges relate to construct definition, confounding factors, and psychometric modeling [13]. But even designing items that foster positive interdependence can be quite tricky. Many “collaborative” tasks can either be solved individually or by dividing the work among the group members rather than through collaboration. For example, the PISA 2015 tasks measuring CPS seem to encourage the test-takers to divide the work with their collaborators [14].

Collaborative learning scholars have emphasized the task design component in contrast to, for example, (over-)scripting student interactions [15]. This approach was central in an online learning and assessment environment called Collaborative Higher-Order Problem Solving (CHOPS) [16]. There, pairs of students work collaboratively to solve math problems built around three item “templates” designed to foster positive interdependence. These templates are described here, using somewhat trivial examples for illustrative purposes:

1. Jigsaw - Students must exchange information to solve the problem, as they have only part of the necessary information. For example, one student might have the length of one side of a rectangle and another student has the length of the adjacent side. Together they are asked to find its area.
2. Joint construction - A correct answer is composed of elements provided by each student that must together satisfy some criteria. For example, each student must provide the length of one side of a rectangle such that its area is 48 units. While there may be multiple solutions, the students must coordinate their responses.
3. Information request - Students have an under-specified problem with limited options to request information to complete the task. The pair must decide together what information is needed and coordinate who should ask for what. For example, the students are asked to determine how long a trip should take and can each request one of the following: the car’s fuel usage, the distance traveled, the car’s average speed, or when the car left its origin.

These templates allow for relatively short-duration CPS items (compared with elaborate scenario tasks). Consequently, many items can be delivered and reliability improved. Item

developers can be trained to adapt many “standard” types of test questions to these templates, but the process is still quite time-consuming.

## **2.2. Large language models**

In recent years, the performance of LLMs, such as OpenAI’s GPT and Meta’s Llama have improved significantly [17]. As a result, these models have been applied in diverse areas such as medicine, computing, basic science, and education [18, 19, 20, 21]. Specifically, GPT-3.5 and GPT-4 have included innovations in bias reduction and complex problem-solving, which are essential for educational applications like content creation, interactive learning, and teaching assistance [22, 23, 24]. Notwithstanding the name “OpenAI”, GPT models are proprietary, potentially expensive, and require users to upload private information to OpenAI servers. Open-source initiatives like Llama and Mistral offer promising alternatives. These models have encouraged an efflorescence of open-source additions, for example other-than-English language capabilities [25].

While LLMs are often remarkably effective at interpreting natural language prompts, higher-quality prompts can yield significantly better outputs [26]. Prompt engineering has emerged as a design problem for refining the content and structure of LLM prompts to optimize for specific tasks [27]. Some prompt engineering best practices involve writing clear, detailed instructions, separating distinct parts of the input, asking the model to adopt a persona, and instructing the LLM to work out the solution rather than immediately constructing the answer [28, 29]. A naming convention has emerged in the literature to describe different prompt variations. Referring to the number of worked examples given to the LLM, *Zero-shot learning* (ZSL) relies solely on the LLM’s pre-trained “knowledge” along with the task description without the use of any worked examples. In contrast, *One-shot learning* (OSL) includes an example in the prompt, and *Few-shot learning* (FSL) includes two or more [30, 31]. There are also variations in the presentation of worked examples. A prompt can include just the correct label or desired response for example. In *Chain-of-thought* (CoT) reasoning [32], however, the prompt demonstrates a multi-step reasoning process, mimicking how a human would approach the problem. These prompting approaches constitute sources of variance that may be important for educational researchers working with LLMs.

## **2.3. Large language models in assessment**

Advancements in LLMs have not gone unnoticed by the measurement field, where they have been considered for item generation, scoring, and parameter calibration [33]. Relatively little research has been conducted on item evaluation using LLMs. Most of this research has focused on automatic evaluation of item difficulty [2]. For instance, researchers used LLM responses to items to evaluate the guessability or the knowledge required to respond to those items [34, 35]. Others have focused on the linguistic features of items [36, 37, 38]. Only a few studies attempted to automatically evaluate items’ content [39], and they generally did not use LLMs for this purpose.

The contribution of LLMs to assessment research and development may be even more pronounced for difficult-to-measure constructs like CPS. Can these models reduce the burden of

new item design? Or will LLM-generated items be disastrous? While LLMs may be able to follow detailed prescriptions for item structure, a more impressive achievement would be understanding the task designer's *intent* more broadly. To that end, a prudent step before engaging an LLM in item generation is to test whether the model has the foundational knowledge to recognize a good CPS item when it sees one.

## 2.4. The current study

In the current study, we sought to examine to what extent LLMs can judge the quality of CHOPS template items for measuring CPS. Given the range of performance demonstrated in the literature, we compared multiple foundational models, prompt strategies, and task types to understand how some approaches may outperform others. This study contributes to the literature in several ways. First, understanding LLMs' ability to evaluate CPS items is a first step in improving item quality and even automatically generating such items. Second, this study is relevant to the measurement field as a whole, as it demonstrates how LLMs deal with complex item evaluation tasks. Finally, by examining different models and prompts we can shed light on the models' respective strengths and limitations, guiding future research in educational technology. In sum, our study aimed to answer the following research questions:

1. To what extent can LLMs evaluate the quality of complex CPS items?
2. To what extent do LLMs' success rates vary by the foundational model, prompting approach, and type of item?

## 3. Methods

### 3.1. Item design

We created a small data set of CPS problems for LLMs to evaluate. The items were designed for two students to solve and are approximately at the level of middle-school math. They use one of three CHOPS templates. We label a CPS task as "good" if it invokes positive interdependence. That is, it requires the participants to work together in a meaningful way to solve the problem. A bad task does not require collaboration or cannot be solved for other reasons. The set contained 21 jigsaw (10 good, 11 bad), 20 joint construction (10 good, 10 bad), and 20 request information (9 good, 11 bad) items, which were either new, adapted from items in CHOPS, or adapted from publicly available large-scale math assessment items like TIMSS and NAEP. Each item was reviewed by at least two team members for clarity, correctness, and content relevance.

Figure 1 shows an example of a joint construction template created based on a TIMSS 2011 item [40]. Version A and B would be shown to the two collaborating students. Since both students can enter values that meet the criterion presented in the item, they do not need to collaborate to solve it, making this a bad example for CPS. A (minimally) good version of this item would require each student to enter one value such that together they meet the criterion. The pair of students must then negotiate a common solution.

Version A:

The minute hand of a clock turns 600 degrees between time T1 and time T2 of the same day. Together with your partner, come up with a possible value for T1 and T2.

■ Enter value for T1 \_\_\_\_\_

◇ Enter value for T2 \_\_\_\_\_

Version B:

The minute hand of a clock turns 600 degrees between time T1 and time T2 of the same day. Together with your partner, come up with a possible value for T1 and T2.

◇ Enter value for T1 \_\_\_\_\_

■ Enter value for T2 \_\_\_\_\_

**Figure 1:** An example of a bad joint construction item. A good variation would keep only the answer input rows preceded with black squares or diamonds but not both.

### 3.2. Pre-prompt design

In this study, we use the term “pre-prompt” when referring to the instructions provided to the LLMs on how to approach the item evaluation, since each “prompt” also includes an item that the LLMs is asked to evaluate. We designed several types of pre-prompts. Initially, we refined one prompt through trial and error with GPT-4 to improve the output. We also designed a pre-prompt following current best practices of asking the LLM to adopt the role of an evaluator and separating its task by first asking it to identify the item type (template) and then make a judgment on collaborative interdependence. Our original prompt was also paired with examples, sometimes limited to pass/fail labels or extended to CoT reasoning. In total, we tested five pre-prompts:

- Zero-shot learning with no examples, prompt refined with GPT-4
- Structured Zero-shot learning following prompt engineering best-practices
- Few-shot learning, original prompt plus one good and bad example from each template (six total); only pass/fail labels were provided
- The same prompt with six CoT examples followed by a verdict
- The same prompt and CoT, except with the verdict given before the reasoning

Below is our ZSL pre-prompt. The CoT pre-prompts with the example items we used for the other pre-prompts, as well as the structured ZSL prompt are available in Appendices A.1 and A.2, respectively.

You will be asked to evaluate one educational exercise for math students working in pairs. The exercise will be presented to you in two parts, the exercise version shown only to Student A (called Version A) and the exercise version as shown only to Student B (Version B). Students A and B are assigned to be partners.

Importantly, Version A and Version B may contain different, complementary information, or the information may be formulated differently. Student A cannot see Version B, and Student B cannot see Version A. The only way they can access the information available to their partner is by communication with each other via text chat. The exercise should require Both Student A and Student B to submit some answers in an answer field or fields.

Your criterion for evaluation of the exercise is whether or not the exercise indeed requires Student A and Student B to collaborate in order to solve the problem. If so, indicate pass. It is not acceptable if Student A and Student B can work separately, independently, and without communicating and still each get the correct answer. In such case, indicate fail. For an exercise to pass, it should be impossible for the students to answer correctly by working alone independently. It is not necessary for you to solve the problem. However, you may describe the solution process in explaining your reasons for your evaluation. When providing your evaluation, please format it as follows:

Verdict: [pass or fail]

Reason: [explanation for verdict]

The following is the exercise you need to evaluate:

### 3.3. Selection of language models

We used six LLMs from three families: GPT-3.5 and GPT-4 from OpenAI, Llama2 and Llama3 from Meta, and Mistral7B and Mixtral8x7B [41, 42]. This selection was designed to explore variance between families as well as within a family, i.e., earlier/later or smaller/larger models. Llama2 and Llama3 come in different sizes; in both cases, we used Q5-quantized versions of the 70 billion (70b) parameter models. Mistral7B is a conventional 7b model. Mixtral8x7b is a Sparse Mixture of Experts (SMoE) architecture with 47b total parameters, but the model uses only 13b at inference time by routing each token to a subset of model components based on the token’s attributes. We used Q8-quantized versions of both Mistral models. The Llama and Mistral models were served locally on a high-performance MacBook Pro with 128GB of RAM.

### 3.4. Procedure

Experimental outputs were collected by an automated script with a browser-based front end. The interface provided for API calls to GPT models and local/cloud-based open-source models. Items and pre-prompts could be selected, and the script would subsequently append the pre-prompts to each item for each call (61 items  $\times$  5 pre-prompts  $\times$  6 models). The outputs of each query were saved for subsequent analyses.

In the analysis stage, LLM outputs were parsed using regular expressions for pass/fail verdicts. All pre-prompts requested verdicts in a specific form, Verdict: Pass/Fail. Model outputs that did not follow this structure were originally parsed as having no verdict. However, further inspection revealed that many model responses contained meaningful evaluations in a different form (e.g., “this exercise meets the criteria”). We therefore wrote a more complex parser to identify relevant phrases. The new parser significantly lowered the no-verdict rates; however, we understand that the parser was still imperfect.

We then compared the results of the parser with our ground-truth labels for each item. The overall agreement is summarized using accuracy (% agreement) and Cohen’s  $\kappa$  [43]. Following

[44] but slightly more conservative at the low end, we interpret Cohen’s  $\kappa$  values  $\leq 0.05$  as poor agreement, 0.06 to 0.20 as slight, 0.21 to 0.40 as fair, 0.41 to 0.60 as moderate, and 0.61 to 0.80 as substantial.

## 4. Results

Table 1 presents the classification performance for all tested models using the ZSL pre-prompt, across all items as well as disaggregated by item type. GPT-4 had the best performance, with an overall moderate agreement level. The three bottom models were barely better than chance (i.e.,  $\kappa$  scores are about zero). Only two open-source foundational models were somewhat comparable to GPT-4: Llama3 and Mixtral8x7b. Overall, Llama3 was better than Mixtral8x7b, but when disaggregated by item type, the results are more complex.

Jigsaw items follow the same pattern as the overall (GPT-4 > Llama3 > Mixtral8x7b). On joint construction items, Llama3 and even Llama2 edge out Mixtral8x7B. However, classifying information request items seems to be the hardest subtask. The highest accuracy, obtained by GPT-4, is 0.63, with a moderate  $\kappa$  of 0.26. Mixtral8x7b slightly beats chance on these items, while Llama3 does worse than chance. In sum, it is possible that to optimize performance using the open-source models, one would do better using Llama3 for jigsaw and joint construction items and Mixtral8x7B for info request items.

**Table 1**

Classification performance (accuracy and  $\kappa$ ) using a common zero-shot prompt for all models. Results are shown for all items as well as for jigsaw (*jig*), joint construction (*jc*), and info request (*ir*) items separately

Model	$Acc_{all}$	$\kappa_{all}$	$Acc_{jig}$	$\kappa_{jig}$	$Acc_{jc}$	$\kappa_{jc}$	$Acc_{ir}$	$\kappa_{ir}$
GPT-4	0.77	0.53	0.86	0.71	0.80	0.60	0.63	0.26
llama3.70B	0.62	0.25	0.81	0.61	0.65	0.30	0.40	-0.14
mixtral8x7b	0.54	0.09	0.62	0.21	0.50	0.00	0.50	0.08
mistral7b	0.51	0.03	0.57	0.10	0.50	0.00	0.45	0.00
llama2.70B	0.51	0.03	0.52	0.00	0.55	0.10	0.45	0.00
GPT-3.5	0.50	0.00	0.53	0.00	0.43	0.00	0.50	0.00

Next, we examined the other pre-prompts to see if they impacted the results. Table 2 includes the classification metrics for the top three performing models, i.e., GPT-4, Llama3, and Mixtral8x7b, across all pre-prompts (the ZSL results from Table 1 are embedded in the first column).

For GPT-4, which had the best overall performance on the task, it is notable that elaboration of the original prompt did not have a positive impact on classification performance and often led to worse performance. The ZSL pre-prompt was as good or better than all others, except CoT prompting for info request items which had identical accuracy and higher  $\kappa$  by about 0.03. However, the difference is probably not of practical significance as the confidence interval around  $\kappa$  is on the order of  $\pm 0.3$ .

While the differences were still small, it does appear that the few-shot prompting improved the results from Llama3 and Mixtral8x7b in a number of prompt-item-type combinations. For



example, CoT prompting improved Mixtral8x7b notably on jigsaw items, while the CoT verdict first improved the joint construction evaluations. Llama3 had more modest gains from these two prompts.

**Table 2**

Accuracy (Cohen’s  $\kappa$ ) for GPT-4, Llama3.70B, and Mixtral8x7B by Pre-prompt Type

GPT-4	GPT-ZSL	Structured ZSL	FSL verdict only	CoT	CoT verdict first
All items	0.77 (0.53)	0.70 (0.40)	0.68 (0.37)	0.75 (0.50)	0.72 (0.43)
Jigsaw	0.86 (0.71)	0.86 (0.71)	0.86 (0.71)	0.81 (0.61)	0.81 (0.61)
Joint construction	0.80 (0.60)	0.70 (0.40)	0.70 (0.40)	0.80 (0.60)	0.75 (0.50)
Info request	0.63 (0.26)	0.53 (0.06)	0.47 (-0.02)	0.63 (0.29)	0.58 (0.19)
Llama3.70B	GPT-ZSL	Structured ZSL	FSL verdict only	CoT	CoT verdict first
All items	0.62 (0.25)	0.59 (0.19)	0.66 (0.32)	0.55 (0.11)	0.64 (0.28)
Jigsaw	0.81 (0.61)	0.76 (0.51)	0.81 (0.61)	0.83 (0.67)	0.71 (0.42)
Joint construction	0.65 (0.30)	0.55 (0.10)	0.65 (0.30)	0.56 (0.12)	0.70 (0.40)
Info request	0.40 (-0.14)	0.45 (0.00)	0.50 (-0.03)	0.44 (-0.11)	0.50 (0.01)
Mixtral8x7B	GPT-ZSL	Structured ZSL	FSL verdict only	CoT	CoT verdict first
All items	0.54 (0.09)	0.54 (0.09)	0.53 (0.06)	0.53 (0.08)	0.53 (0.07)
Jigsaw	0.62 (0.21)	0.67 (0.31)	0.62 (0.22)	0.76 (0.52)	0.52 (0.02)
Joint construction	0.50 (0.00)	0.55 (0.10)	0.50 (0.00)	0.47 (-0.03)	0.63 (0.23)
Info request	0.50 (0.08)	0.40 (-0.10)	0.45 (0.00)	0.35 (-0.25)	0.45 (-0.04)

The above analysis is perhaps too fine, slicing by model, prompt, and item type. To understand if different prompts are generally more suitable to different item types, we average over the top three models. These results are shown in Table 3. Indeed, after averaging, it remains the case that the best overall prompt is not the best prompt for each item type. Notably, the classification of info request items is, at best, barely better than chance.

**Table 3**

Best preprompt (using Cohen’s  $\kappa$ ) Overall and by Item Type. The highest values in each column are bolded.

Prompt	$\kappa_{all}$	$\kappa_{jig}$	$\kappa_{jc}$	$\kappa_{ir}$
GPT-ZSL	<b>0.29</b>	0.51	0.30	<b>0.06</b>
Structured ZSL	0.22	0.51	0.20	-0.02
FSL verdict only	0.24	0.51	0.23	-0.01
CoT	0.22	<b>0.54</b>	0.22	-0.03
CoT verdict first	0.26	0.35	<b>0.38</b>	<b>0.06</b>

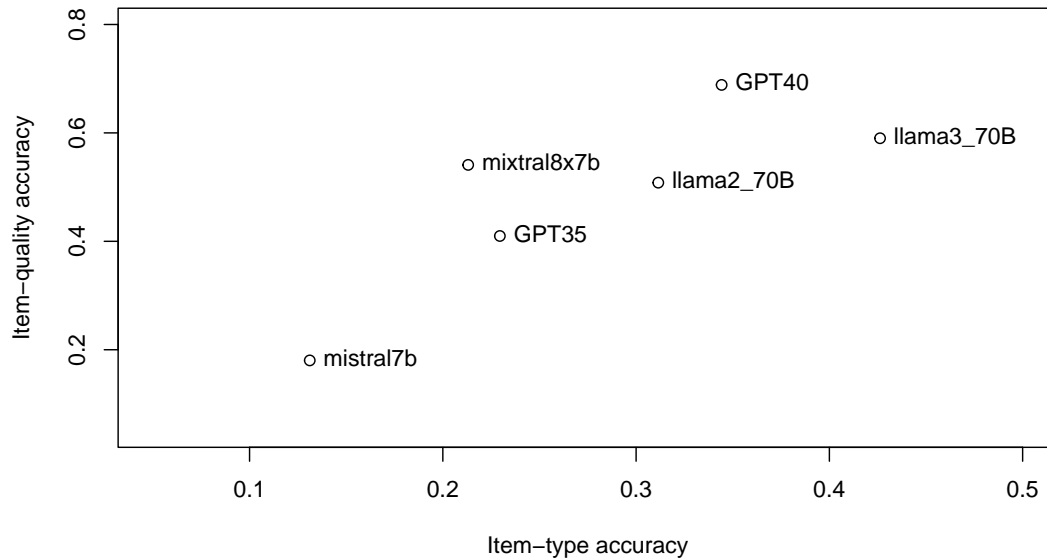
It appears to be the case that jigsaw classification is the most successful, followed by joint construction and information request. A high-level summary confirming this finding using accuracy scores averaged over pre-prompts for each model is shown in Table 4. Note that these are not the best results for each model.



**Table 4**

Average Accuracy by Item Type and Model

Model	Jigsaw	Joint construction	Information request
GPT-4	0.84	0.75	0.54
llama3.70B	0.76	0.61	0.46
mistral8x7b	0.64	0.52	0.43

**Figure 2:** Relationship between item-type classification success and item quality evaluation using the structured ZSL prompt.

As an exploratory step, we were interested in whether the models were able to classify items into the correct types, the first sub-task using the structured ZSL approach. Base rate classification accuracy for item types could be expected at 0.33, and actual results ranged from 0.13 to 0.43. Striking, however, is the relationship between accuracy in classifying the item type (template) and accuracy in evaluating the items (see Figure 2). The highly apparent correlation ( $r = 0.80$ ) suggests that better models in one task can do the other better as well. Interestingly, when it came to type classification, Llama3 was actually the best performing model.

## 5. Discussion

The purpose of our study was to test the feasibility of LLMs for evaluating items measuring CPS. We also wanted to see if different models, pre-prompts, or item types affect the results.

Understanding these issues may contribute to research on how LLMs interact with complex tasks and to future item design in practice. According to our findings, only three of the tested models did better than chance, with GPT-4 outperforming the other models in almost all cases. Between the open-access models, different models did better on different item types, suggesting that users should consider the task type when choosing the best model. Given GPT-4's success relative to other models in various tasks [45, 46], including tasks related to item generation [47], this result is unsurprising. However, even GPT-4 reached only moderate levels of agreement in most cases. Others have also found that LLMs struggle with evaluative tasks [48], suggesting directions for future LLM developments.

Contrary to existing findings [49], elaborate pre-prompting rarely improved on the basic ZSL pre-prompt. It is possible that the examples were confusing or focused the LLMs on the specific cases rather than the general idea. We intend to examine this issue in the future. We also found that some item types were easier for the LLMs to judge than others. All models generally did best with the jigsaw items followed by the joint construction items and the information request items. We are unaware of existing research comparing LLMs' ability to evaluate different types of interdependent tasks, and this might also be a fruitful direction for future work.

This study has several limitations. First, our basic ZSL pre-prompt was refined using GPT-4, perhaps contributing to its success. Since GPT-4 seems to outperform other models in a variety of complex tasks, we believe this effect is likely small. Second, to enhance the study's generalizability, more items, constructs, models, and pre-prompts should be tested. Finally, we could only examine the final verdict of the models and not their reasoning. Qualitative analysis of the LLMs' outputs is planned and could reveal the reasons for their disagreements with humans.

In conclusion, when evaluating the quality of CPS items, existing LLMs have only moderate levels of agreement with humans at best. Adding more information beyond ZSL pre-prompts does not improve this by much. However, different models and pre-prompts perform better when evaluating different item types. Therefore, more work on the models or on prompting strategies is required before LLMs can be reliably used for evaluating items measuring CPS and, likely, similarly complex constructs.

## References

- [1] J. A. Rios, G. Ling, R. Pugh, D. Becker, A. Bacall, Identifying critical 21st-century skills for workplace success: A content analysis of job advertisements, *Educational Researcher* 49 (2020) 80–89.
- [2] L. Benedetto, P. Cremonesi, A. Caines, P. Buttery, A. Cappelli, A. Giussani, R. Turrin, A survey on recent approaches to question difficulty estimation from text, *ACM Computing Surveys* 55 (2023) 1–37. doi:10.1145/3556538.
- [3] J. Burrus, T. Jackson, N. Xi, J. Steinberg, Identifying the most important 21st century workforce competencies: An analysis of the occupational information network (o\*net), *ETS Research Report Series 2013* (2013) i–55.
- [4] L. Darling-Hammond, J. Herman, J. Pellegrino, J. Abedi, J. L. Aber, E. Baker, R. Bennett, E. Gordon, E. Haertel, K. Hakuta, et al., Criteria for high-quality assessment, *Stanford Center for Opportunity Policy in Education* 2 (2013) 171–192.
- [5] S. M. Fiore, A. Graesser, S. Greiff, P. Griffin, B. Gong, P. Kyllonen, C. Massey, H. O’Neil, J. Pellegrino, R. Rothman, et al., Collaborative problem solving: Considerations for the national assessment of educational progress (2017).
- [6] R. M. Gillies, Cooperative learning: Review of research and practice, *Australian Journal of Teacher Education (Online)* 41 (2016) 39–54.
- [7] P. A. Kirschner, F. Kirschner, J. Janssen, The collaboration principle in multimedia learning, *The Cambridge handbook of multimedia learning* 2 (2014) 547–575.
- [8] D. Mullins, N. Rummel, H. Spada, Are two heads always better than one? differential effects of collaboration on students’ computer-supported learning in mathematics, *International Journal of Computer-Supported Collaborative Learning* 6 (2011) 421–443.
- [9] D. W. Johnson, R. T. Johnson, An educational psychology success story: Social interdependence theory and cooperative learning, *Educational researcher* 38 (2009) 365–379.
- [10] R. E. Slavin, Research on cooperative learning and achievement: What we know, what we need to know, *Contemporary educational psychology* 21 (1996) 43–69.
- [11] B. M. Stecher, L. S. Hamilton, Measuring Hard-to-Measure Student Competencies: A Research and Development Plan. Research Report., ERIC, 2014.
- [12] A. C. Graesser, P. W. Foltz, Y. Rosen, D. W. Shaffer, C. Forsyth, M.-L. Germany, Challenges of assessing collaborative problem solving, *Assessment and teaching of 21st century skills: Research and applications* (2018) 75–91.
- [13] J. Hao, L. Liu, A. A. von Davier, P. C. Kyllonen, Initial steps towards a standardized assessment for collaborative problem solving (cps): Practical challenges and strategies, *Innovative assessment of collaboration* (2017) 135–156.
- [14] Pisa 2015 collaborative problem solving, <https://www.oecd.org/pisa/innovation/collaborative-problem-solving/>, Accessed: 2024-05-10.
- [15] P. Dillenbourg, Over-scripting cscl: The risks of blending collaborative learning with instructional design., *Three worlds of CSCL. Can we support CSCL?* (2002) 61–91.
- [16] Y. Bergner, Y. Wang, Mathchops: A platform for developing collaborative higher order problem solving in mathematics, in: *Proceedings of the 16th International Conference on Computer-Supported Collaborative Learning-CSCL 2023*, pp. 51-58, International Society of the Learning Sciences, 2023.

- [17] K. I. Roumeliotis, N. D. Tselikas, Chatgpt and open-ai models: A preliminary review, *Future Internet* 15 (2023) 192.
- [18] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, D. S. W. Ting, Large language models in medicine, *Nature medicine* 29 (2023) 1930–1940.
- [19] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günemann, E. Hüllermeier, et al., Chatgpt for good? on opportunities and challenges of large language models for education, *Learning and individual differences* 103 (2023) 102274.
- [20] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, P. Schwaller, Chemcrow: Augmenting large-language models with chemistry tools, *arXiv preprint arXiv:2304.05376* (2023).
- [21] F. F. Xu, U. Alon, G. Neubig, V. J. Hellendoorn, A systematic evaluation of large language models of code, in: *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming, 2022*, pp. 1–10.
- [22] D. Kovačević, Use of chatgpt in esp teaching process, in: *2023 22nd International Symposium INFOTEH-JAHORINA (INFOTEH)*, IEEE, 2023, pp. 1–5.
- [23] J. Rudolph, S. Tan, S. Tan, Chatgpt: Bullshit spewer or the end of traditional assessments in higher education?, *Journal of applied learning and teaching* 6 (2023) 342–363.
- [24] A. Tlili, B. Shehata, M. A. Adarkwah, A. Bozkurt, D. T. Hickey, R. Huang, B. Agyemang, What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education, *Smart Learning Environments* 10 (2023) 15.
- [25] A. Balachandran, Tamil-llama: A new tamil language model based on llama 2, *arXiv preprint arXiv:2311.05845* (2023).
- [26] Z. Lin, How to write effective prompts for large language models, *Nature Human Behaviour* (2024) 1–5.
- [27] T. Sorensen, J. Robinson, C. M. Rytting, A. G. Shaw, K. J. Rogers, A. P. Delorey, M. Khalil, N. Fulda, D. Wingate, An information-theoretic approach to prompt engineering without ground truth labels, *arXiv preprint arXiv:2203.11364* (2022).
- [28] Best practices for prompt engineering with the openai api, [https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api#h\\_eae065300d](https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api#h_eae065300d), 2024. Accessed: 2024-05-04.
- [29] Prompt engineering, <https://platform.openai.com/docs/guides/prompt-engineering>, n.d. Accessed: 2024-05-04.
- [30] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. 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) 1877–1901.
- [31] J. Chen, Y. Geng, Z. Chen, J. Z. Pan, Y. He, W. Zhang, I. Horrocks, H. Chen, Zero-shot and few-shot learning with knowledge graphs: A comprehensive survey, *Proceedings of the IEEE* (2023).
- [32] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, *Advances in neural information processing systems* 35 (2022) 24824–24837.
- [33] J. Hao, A. A. von Davier, V. Yaneva, S. Lottridge, M. von Davier, D. J. Harris, Transforming assessment: The impacts and implications of large language models and generative ai,

Educational Measurement: Issues and Practice (2024).

- [34] H. Maeda, Field-testing multiple-choice questions with ai examinees (2024).
- [35] A. Säuberli, Automatic Generation and Evaluation of Multiple-Choice Reading Comprehension Items with Large Language Models, Ph.D. thesis, University of Zurich, 2023.
- [36] R. Rodriguez-Torrealba, E. Garcia-Lopez, A. Garcia-Cabot, End-to-end generation of multiple-choice questions using text-to-text transfer transformer models, *Expert Systems with Applications* 208 (2022) 118258.
- [37] V. Raina, M. Gales, Multiple-choice question generation: Towards an automated assessment framework, *arXiv preprint arXiv:2209.11830* (2022).
- [38] M. J. Gierl, H. Lai, V. Tanygin, *Advanced methods in automatic item generation*, Routledge, 2021. doi:10.4324/9781003025634.
- [39] R. Meissner, D. Jenatschke, A. Thor, Evaluation of approaches for automatic e-assessment item annotation with levels of bloom's taxonomy, in: *International Symposium on Emerging Technologies for Education*, Springer, 2020, pp. 57–69.
- [40] International Association for the Evaluation of Educational Achievement (IEA), TIMSS 2011 Assessment, TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College, Chestnut Hill, MA and International Association for the Evaluation of Educational Achievement (IEA), IEA Secretariat, Amsterdam, the Netherlands., 2013.
- [41] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, *arXiv preprint arXiv:2310.06825* (2023).
- [42] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al., Mixtral of experts, *arXiv preprint arXiv:2401.04088* (2024).
- [43] J. Cohen, Coefficient of agreement for nominal scales, *Educational and psychological measurement* 20 (1960).
- [44] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, *biometrics* (1977) 159–174.
- [45] A. Borji, M. Mohammadian, Battle of the wordsmiths: Comparing chatgpt, gpt-4, claude, and bard, GPT-4, Claude, and Bard (June 12, 2023) (2023).
- [46] L. Martin, N. Whitehouse, S. Yiu, L. Catterson, R. Perera, Better call gpt, comparing large language models against lawyers, *arXiv preprint arXiv:2401.16212* (2024).
- [47] A. Säuberli, S. Clemenide, Automatic generation and evaluation of reading comprehension test items with large language models, *arXiv preprint arXiv:2404.07720* (2024).
- [48] J. Steiss, T. Tate, S. Graham, J. Cruz, M. Hebert, J. Wang, Y. Moon, W. Tseng, M. Warschauer, C. B. Olson, Comparing the quality of human and chatgpt feedback of students' writing, *Learning and Instruction* 91 (2024) 101894.
- [49] B. Chen, Z. Zhang, N. Langrené, S. Zhu, Unleashing the potential of prompt engineering in large language models: a comprehensive review, *arXiv preprint arXiv:2310.14735* (2023).

## A. Full text of prompts

### A.1. Chain-of-Thought

You will be asked to evaluate one educational exercise for math students working in pairs. The exercise will be presented to you in two parts, the exercise version shown only to Student A (called Version A) and the exercise version as shown only to Student B (Version B). Students A and B are assigned to be partners. Importantly, Version A and Version B may contain different, complementary information, or the information may be formulated differently. Student A cannot see Version B, and Student B cannot see Version A. The only way they can access the information available to their partner is by communication with each other via text chat. The exercise should require Both Student A and Student B to submit some answers in an answer field or fields.

Your criterion for evaluation of the exercise is whether or not the exercise indeed requires Student A and Student B to collaborate in order to solve the problem. If so, indicate pass. It is not acceptable if Student A and Student B can work separately, independently, and without communicating and still each get the correct answer. In such case, indicate fail. For an exercise to pass, it should be impossible for the students to answer correctly by working alone independently. It is not necessary for you to solve the problem. However, you may describe the solution process in explaining your reasons for your evaluation. When providing your evaluation, please format it as follows:

Verdict: [pass or fail]

Reason: [explanation for verdict]

##The following is an example exercises with suitable response:

#Example prompt

Version A: A factory produces 100,000 batteries each day. A sample of 200 batteries is drawn from today's production line, and 2 batteries fail the quality test. What is the best estimate for the total number of faulty batteries produced today?

Version B: A factory produces 100,000 batteries each day. A sample of 200 batteries is drawn from today's production line, and 2 batteries fail the quality test. What is the best estimate for the total number of faulty batteries produced today?

#Example response

To estimate the total number of faulty batteries produced, one needs to know the total daily production, the size of the test sample, and the number of failed batteries in the test sample. Both Student A and Student B have the complete information needed to solve the problem and thus can in principle solve the problem without collaborating with one another.

Verdict: Fail

#Example prompt

Version A: A factory produces batteries each day. A sample of 200 batteries is drawn from today's production line, and 2 batteries fail the quality test. What is the best estimate for the total number of faulty batteries produced today?

Version B: A factory produces 100,000 batteries each day. A sample of batteries is drawn from today's production line, and 2 batteries fail the quality test. What is the best estimate for the total number of faulty batteries produced today?

#Example response

To estimate the total number of faulty batteries produced, one needs to know the total daily production, the size of the test sample, and the number of failed batteries in the test sample. Student A has the sample size but does not have the total number produced, while Student B knows the total number of batteries produced but does not know the size of the sample that was tested. The collaborating students need to communicate this information to each other to estimate the total number of faulty batteries produced today. Thus, this exercise meets the requirement that it can only be solved if Student A and Student B share information with each other.

Verdict: Pass

#Example prompt

Version A:

The minute hand of a clock turns 600 degrees between time T1 and time T2 of the same day. Together with your partner, come up with a possible value for T1 and T2.

Enter value for T1:

Enter value for T2:

Version B:

The minute hand of a clock turns 600 degrees between time T1 and time T2 of the same day. Together with your partner, come up with a possible value for T1 and T2.

Enter value for T1:

Enter value for T2:

#Example response



There is an infinite number of possible solutions to the posed problem. Each student is provided with the ability to provide a complete solution to the problem. Thus, it is possible for each student to answer correctly on their own without coordinating with their partner.

Verdict: Fail

#Example prompt

Version A:

The minute hand of a clock turns 600 degrees between time T1 and time T2 of the same day. Together with your partner, come up with a possible value for T1 and T2.

Enter value for T1:

Version B:

The minute hand of a clock turns 600 degrees between time T1 and time T2 of the same day. Together with your partner, come up with a possible value for T1 and T2.

Enter value for T2:

#Example response

Each student is provided with the ability to answer one of two necessary parts of the solution. Moreover, the two parts must together compose a correct solution. Although there is an infinite number of possible solutions to the posed problem, neither student can answer correctly on their own without coordinating with their partner.

Verdict: Pass

#Example prompt

Version A: In a school fund-raiser, students in class A and class B sold boxes of cookies. What was the average number (arithmetic mean) of boxes of cookies sold by all students in both classes?

To answer this question, you and your partner may each make TWO selections from the following list of values. After you submit your selection, the values you selected will be revealed to you. Use this information to provide your answer in the box below.

- A. Average number of boxes of cookies sold in class A
- B. Total number of boxes of cookies sold in class A
- C. Average number of boxes of cookies sold in class B
- D. Total number of cookies per box
- E. Total number of students in class A

Version B: In a school fund-raiser, students in class A and class B sold boxes of cookies. What was the average number (arithmetic mean) of boxes of cookies sold by all students in both classes?

To answer this question, you and your partner may each make TWO selections from the following list of values. After you submit your selection, the values you selected will be revealed to you. Use this information to provide your answer in the box below.

- A. Average number of boxes of cookies sold in class A
- B. Total number of boxes of cookies sold in class A
- C. Average number of boxes of cookies sold in class B
- D. Total number of cookies per box
- E. Total number of students in class A

#Example response

Critical pieces of information necessary for solving the problem (such as the total number of students in both classes or the total number of boxes sold in class B) are either missing or inadequately defined in the options available to the students. Therefore, the task is unsolvable with the provided selections, even if students work together to combine their available information. The exercise does not meet the criteria for a solvable and collaborative educational exercise.

Verdict: Fail

#Example prompt

Version A: In a school fund-raiser, students in class A and class B sold boxes of cookies. What was the average number (arithmetic mean) of boxes of cookies sold by all students in both classes?

To answer this question, you and your partner may each make TWO selections from the following list of values. After you submit your selection, the values you selected will be revealed to you. Use this information to provide your answer in the box below.

- A. Average number of boxes of cookies sold in class A
- B. Total number of boxes of cookies sold in class A
- C. Average number of boxes of cookies sold in class B
- D. Total number of boxes of cookies sold in class B
- E. Total number of cookies per box
- F. Total number of students in class A
- G. Total number of students in class B

Version B: In a school fund-raiser, students in class A and class B sold boxes of cookies. What was the average number (arithmetic mean) of boxes of cookies sold by all students in both classes?

To answer this question, you and your partner may each make TWO selections from the following list of values. After you submit your selection, the values you selected will be revealed to you. Use this information to provide your answer in the box below.

- A. Average number of boxes of cookies sold in class A
- B. Total number of boxes of cookies sold in class A
- C. Average number of boxes of cookies sold in class B
- D. Total number of boxes of cookies sold in class B
- E. Total number of cookies per box
- F. Total number of students in class A
- G. Total number of students in class B

#Example response

To calculate the overall average number of boxes sold by students in both classes, students will need at least four pieces of information from the options provided. For instance, one student might choose the total number of boxes sold in class A and the total number of students in class A, while the other selects the equivalent information for class B. Alternatively, they could choose average numbers and total students in each class. However, each student has the ability to select only two pieces of information. Without sharing this information, neither student can independently calculate the overall average, fulfilling the requirement for collaboration.

Verdict: Pass

The following is the exercise you need to evaluate:

## A.2. Structured Zero-shot

Your Role: Collaboration evaluator for math exercises

Objective: You need to evaluate collaborative math exercises provided for two students who are solving the exercises together. The goal of this evaluation is to determine whether the exercises require genuine collaboration between the partners to solve.

Exercise overview: Each exercise will be presented to you in two parts, Version A, accessible only to Student A, and Version B, accessible only to Student B. Students A and B are assigned to be partners.

Types of collaborative exercises:

1. Jigsaw (the pair of students are provided different or complementary information that needs to be shared to arrive at the solution)
2. Joint construction (the pair of students are provided the same information but need to solve and respond with different parts of the solution)
3. Info request (the students may or may not receive different information, but they will need to collaborate to identify two pieces of information they can request to solve the exercise)

Thus, Version A and Version B may contain different or complementary information, the information may be formulated differently, or the response options provided to each student may be different. Images or figures provided are summarized in text within square brackets. Student A cannot see Version B, and Student B cannot see Version A. The only way they can access the information available to their partner is by communication with each other via text chat. The exercise should require both Student A and Student B to submit some answer(s).

Evaluation Criteria:

1. Communication Necessity: Is communication between Student A and Student B essential for completing the exercise?
2. Solution Process: Can the problem only be solved through the combined efforts and information of both students?

It is not necessary for you to solve the problem. However, you may describe the solution process in explaining your reasons for your evaluation.

Evaluation format: When providing your evaluation, please format it as follows: Verdict: [pass or fail]

Type: [Jigsaw, Joint Construction, Info Request, NA (if fail), Other (if pass but does not fit any of the types)]

Reason: [explanation for verdict]