A case study using Large Language Models to generate metadata for math questions

Katie Bainbridge [1], Candace Walkington[2], Armon Ibrahim [1], Iris Zhong [1], Debshila Basu Mallick [1], Julianna Washington[2], Rich Baraniuk [1]
[1] *OpenStax, Rice University, Houston, Texas, United States*
[2] *Southern Methodist University, Dallas, Texas, United States*

Abstract
Creating labels for assessment items, such as concept used, difficulty, or vocabulary used, can improve the quality and depth of research insights as well as targeting the right kinds of questions for students depending on their needs. However, traditional processes for metadata tagging are resource intensive in terms of labor, time, and cost, and these metadata become quickly outdated with any changes to the question content. Given thoughtful prompts, Large Language Models (LLMs) like GPT 3 and 4 can efficiently automate generation of assessment metadata and can help scale the process for larger volumes of questions as well as address any updates to question content that would otherwise have been tedious to reanalyze. With a human subject matter expert in-the-loop, recall and precision was analyzed for LLM generated tags for two metadata variables: problem context and math vocabulary. It is concluded that LLMs like GPT 3.5 and 4 are highly reliable at generating assessment metadata, and recommendations for others hoping to apply the technology to their own assessment items are made.

Keywords
Large Language Models, Assessments, Metadata, Human-in-the-loop

1. Introduction

Learning Sciences research often requires extensive metadata for assessment items in order to make meaningful insights about student learning. For example, if we know what concept each question targets, we can track a student's competency in that concept over the course of the school year [1]. We aimed to examine the role reading comprehension plays in math achievement [2,3], specifically in middle school algebra. This research direction required that we create a new suite of metadata for our assessment bank that gauged the various factors that might affect a student's reading comprehension. Some of these variables, such as word count, are easy to automate, but others, such as whether a question is set in a real world context like a baseball game, must be done manually. Manual metadata generation poses a number of logistical problems; it takes a lot of time and resources, it is tedious, and ones' efforts are quickly made irrelevant once any of the source content is updated or edited. Faced with multiple variables that required manual coding, we were motivated to find a way to automate as many as possible.

One such example is *question context*. Math problems that are contextualized in a real-world context require higher reading comprehension skills than math problems that are purely symbolic. A student who struggles with reading is much more likely to have their true understanding of the underlying math concepts obscured by their poor reading comprehension on a problem like this:

AIED2023, Empowering Education with LLMs - the Next-Gen Interface and Content Generation, July 7,2023
EMAIL: katie@openstax.org (A. 1); cwalkington@mail.smu.edu (A. 2); aji1@rice.edu (A. 3) iz5@rice.edu (A 4); debshila@rice.edu (A 5); richb@rice.edu (A 7)
ORCID: 0000-0003-2224-5140 (A. 1); 0000-0002-2338-8760 (A. 2); 0000-0002-0955-5884; (A. 4); 0000-0002-0597-3528 (A 5); 0000-0002-0721-8999 (A7)

"These equations represent the number of bacteria in four different dishes as a function of time, t , in days. Which equation represents the population with the greatest growth factor?"

Than on a problem like this:

"What is the solution to 4(y - 3) + 19 = 8(2y + 3) + 7?"

The former requires the student to have familiarity with a greater number of vocabulary words, both math vocabulary like "function" and "equation", but also real world vocabulary such as "bacteria"[4,5]. It also expects the student to be able to generate an accurate problem model of the relationships between the terms [6,7]. If a student struggles to read the problem or translate the problem into an accurate problem model [8], or if they are spending too much cognitive load deciphering the unfamiliar vocabulary [9], they may not be able to demonstrate their true understanding of the underlying concept: growth factors.

In order to research the role reading plays in math success, and in order to potentially intervene with appropriate support when students are struggling due to reading rather than computation, it is helpful to know whether a given question is contextualized in a real-world situation or if it is purely symbolic. Making such a judgment manually is relatively easy, but doing so for hundreds of questions takes time.

Luckily, LLMs like GPT 3.5 and GPT 4 are well-suited to making judgments like this. This paper documents the process we used to generate metadata using GPT 3.5 and GPT 4 for two variables: Context and Math Vocabulary. As part of our overall methodology, we include the prompts used, the errors generated, revisions to the prompts, our QA process, and accuracy, reliability, and precision analyses comparing the AI generated tags to those created by a human coder. We conclude with lessons learned and recommendations for others hoping to use LLMs to generate metadata for math content as well as other domains.

1.2 General Methods

We integrated OpenAI's GPT 3.5 with Google Sheets using an extension available at GPTforWork.com [10]. In a column adjacent to the question text we queried using the formula =GPT(quesiton_text, "text of our request to GPT"). The formula could then be quickly repeated across rows for the whole column. Initial queries underwent a QA process, comparing a sample of GPT tags to the manual tags created by a human subject matter expert (SME). Recurrent errors were identified, and the prompts were iteratively refined to address the errors. For example, GPT model had a bad habit of trying to answer the math question in addition to responding to the prompt, so subsequent queries instructed it not to solve the math problem.

Once we were satisfied with the prompt, we cleaned the data so that answer formatting was consistent. GPT would sporadically put a period at the end of a response, at times it would explain its reasoning after giving a response, and despite instructions it would still occasionally include the solution to the math problem. These inconsistencies were resolved and unrequested additions were removed so that responses could be compared statistically.

Once responses were cleaned, that LLM generated metadata were compared against the SME generated metadata and evaluated with recall and precision measures.

2. Context Extraction Methods

Whether a math question is contextualized with a real-world application can greatly increase the role that reading plays in a students' ability to solve a math problem. We labeled questions like this "Real-World". Questions that use purely symbolic math were labeled "Symbolic". A third category that was not discussed in the introduction was labeled "School

Math". These questions are symbolic math questions that are contextualized with a hypothetical student who is trying to solve the math problem. They can be used to assess metacognitive knowledge of math procedures, e.g. *"What should Mai do next?"*, or to assess a students' ability to identify errors or misconceptions, e.g. *"Noah is solving an equation, and one of his moves is unacceptable. Here are the moves he made:"*. These require less reading comprehension than "Real-World" questions, but more reading comprehension than "Symbolic" questions.

We had a total of 339 quetions in our assessment bank. 109 were identified by SME as having a real world context, 224 were identified as being purely symbolic, and 6 were identified as "School Math". In our initial experimentation with this idea using the Chat GPT interface, we'd already established that GPT 3.5 had trouble distinguishing School Math from Symbolic questions. Thus, the first prompt for GPT 3.5 was "Is the math problem set in a real-world context, or is it symbolic math? If it is set in a real-world context, say Real-World, if it does not have a real world context say Symbolic". Upon inspection, GPT 3.5 routinely made the following errors:

1. Solving the math question in addition to answering the prompt.
2. Considering a graph to be a real-world context, even if the math was symbolic

Our prompt was updated to: *"Is the math problem set in a real-world context, or is it symbolic math? If it is contextualized in a real-world setting, say Real-World, if it is not applying math in a real-world environment, say Symbolic. Do not solve the math problem."*. The types of errors made by this prompt were artifacts of our question format; question images and answer options for multiple choice questions were not included in the question text given to GPT. For problems that relied heavily on either images or answer option text (e.g. "What is true about the following diagram?" or "Which rule can describe the table below"), GPT 3.5 would either respond that the question was in both categories (n=9) or it would say that it needed more information (n=11). A total of 20 questions out of 339 questions were removed from the analysis for this variable.

We then created a different prompt on just the symbolic questions to separate the School Math questions. This prompt read: *"Does the math question contain a person's name? If the answer is yes, respond with School Math, if the answer is no respond with Symbolic. Do not solve the math problem"*. This prompt did not capture our sampled cases, so we revised it to: *"Is this math question about a hypothetical person trying to solve a symbolic math problem? Does the question text contain a person's name? If the answer to either is yes, respond with School Math, if the answer is no respond with Symbolic. Do not solve the math problem"*. The "Real-World" labels for the remaining questions were combined with the resulting list.

2.1. Context Evaluation Results

Manual Tags and GPT 3.5 tags generated on our 319 items from our assessment bank were compared on recall (0.92), and precision (0.92). Results can be seen in Table 1.

Table 1
Evaluation results for the GPT 3.5 generated context metadata with the SME created metadata.

Test	Result
Recall	0.92
Precision	0.92

3. Math Vocabulary Extraction Methods

We identified "math vocabulary" as a second variable that would be a good candidate for using GPT rather than a human coder. Words like "linear" and "quadratic" are key to understanding Algebra questions; if a student needs support on vocabulary, these words pose the biggest barrier to their ability to demonstrate understanding. Real-world vocabulary, like the "bacteria" example described in the introduction, is a separate variable.

The first prompt for GPT 3.5 was framed as *"List the math vocabulary words in the question text. Do not answer the math question."*. The errors that resulted fell into two categories. Primarily, GPT 3.5 would sometimes identify the mathematical expressions in the question as math vocabulary. For example, for the question text *"Which equation is equivalent to the equation 6x + 9 = 12?",* it provided *"Equation, 6x + 9 = 12"* in response to our prompt. Secondly, GPT would only sometimes identify *"graph"* as math vocabulary, whereas the human coder always considered *"graph"* to be math vocabulary. For multi-word phrases, such as *"linear relationship"*, GPT will sometimes return the whole phrase but at other times will only return part of the phrase, such as *"linear"*; however, the human coder could occasionally be inconsistent about this as well.

At times GPT 3.5 could be considered *more* accurate than the human coder. For example, the human coder did not consider "*data"* to be math vocabulary, whereas GPT did. This can be seen as a benefit to using LLMs for this task; the risks of a "false positive" in this case are minimal. If GPT identifies additional words for which a student might need support, that is a benefit rather than a drawback.

Our second, revised prompt read *"Excluding numbers, variables, mathematical expressions and equations, list only the math vocab words/phrases in the question text. Do not answer the question"*. This phrasing successfully captured multi-word phrases in the sample we reviewed, and the instances of numbers and expressions being included were reduced (but not eliminated). The new phrasing did not reduce the instances of inconsistently considering words like "graph" to be math vocabulary. As we did not know why this inconsistency happened, we could not think of a way to address it in our prompt.

The process was repeated using GPT 4, starting with the prompt *"Excluding numbers, variables, mathematical expressions and equations, list only the math vocab words/phrases in the question text. Do not answer the question"*. The resulting errors suggested that GPT 4 was much more liberal in what it considered to be "mathematical vocabulary". It included most of the real-world vocabulary (e.g. *"softball team"* and *"landscaping company"*) in its response to the prompt. We changed the prompt to say *""List the math vocabulary words in the selected question text. Only include mathematical vocabulary, do not include vocabulary without a mathematical definition. Do not answer the math question."*. The removed many cases of including real-world context vocabulary, but many still remained. Some, such as "nickles", one could make an argument for having a "mathematical definition"; however for others, such as "pretzles", it is harder to make an argument for why GPT 4 considered the term to be mathematical vocabulary.

The problems seen in our previous use of GPT 3.5, such as inconsistent formatting for the response, and including expressions and variables in the response, persisted with GPT4. For our cleaning process we made a second column and, referencing the list of vocabulary GPT 4 had just created, provided this prompt: *"Reformat the text in this cell as a list separated by commas. Remove all periods, mathematical expressions, solitary letters, and variables representing numbers".*

GPT 4 also introduced a new problem, in that in some cases it considered the text from the prompt in its response. This led to the inclusion of the phrase "mathematical vocabulary" in

dozens of responses, despite the phrase never appearing in the question bank text. Language from the prompt was removed manually before analysis.

3.1 Math Vocabulary Evaluation Results

In cases like this, where the consequences of the AI identifying a case where a human did not (false positive) are null, a Recall test would be the most accurate. The Recall analysis was done by dividing the number of matches that GPT made (true positives) by the total number of words identified by the human coder for each question. This recall score was averaged across questions. It resulted in a recall of 0.75; GPT 3.5 successfully identified the vocabulary the human coder did in 75% of cases. (See Table 2)

Table 2

Evaluation results for the GPT 3.5 vs SME generated math vocabulary metadata

Test	Result
Recall	0.75
Precision	0.63

Precision (true positives divided by total words identified by GPT) was lower than recall (0.63), suggesting that false positives were numerous enough to negatively impact precision. However, as discussed previously, the consequences of false positives are minimal (and may even be beneficial), so this result should carry less weight than the rate of recall. Interpreting these results is not as straightforward as the interpretation for context, as each response could have multiple potential matches. In many ways, if GPT looks at a math question and identifies 4 vocabulary words, where a human saw 5 vocabulary words, that is still a relatively successful application of this technology despite the slight decrease in recall it represents. The question is now whether a recall rate of 0.75 is high enough to warrant relying on AI-generated tags in place of human-generated ones, although it must be noted that tags that have a single categorical response will naturally result in a higher recall rate.

We repeated our math vocabulary tagging process using GPT 4 to see if it improved recall (see Table 3). GPT 4 was indeed better at identifying vocabulary, with a recall rate of 0.82, indicating that GPT 4 agreed with the human coder on 82% of the identified vocabulary words. Precision only moderately increased (from 0.63 to 0.66), indicating that false positives are still high.

Table 2

Evaluation results for the GPT 4 vs SME generated math vocabulary metadata

Test	Result
Recall	0.75
Precision	0.63

These recall results get closer to the margin of error we would expect between two human coders, especially for a subjective, multi-class labeling task [11,12]. Inspection of the false positives made by GPT 4 include things like a greater-than-or-less-than symbol (≥), which is indeed a type of math vocabulary for which a student may need support. Many of the false positives are cases of real-world conext vocabulary being included despite our attempts to remove them with our prompt experimentation. We intend to expand our metadata tags to include this type of vocabulary as a separate variable. A promising next step would be to repeat this process for the real-world vocabulary, then undergo a cleaning process that removes redundant words from the math vocabulary generated by GPT 4

4. Discussion

In the comparison between GPT 3.5 and a human coder on creating metadata for math, specifically, algebra questions, the results were mixed. Context was a success, with GPT correctly labeling the

problems with 92% recall and precision. Vocabulary was less successful, with a 75% recall rate. GPT 4 was more successful at the vocabulary task, raising the recall to 82%. Overall we conclude that using AI to generate metadata for algebra questions shows promise and should be explored further. For many applications, the time, cost, and labor saved is worth the modest decreases to accuracy, particularly for cases where false positives carry few consequences.

4.1 Recommendations

We learned much in the process of automating the metadata generation of math assessments that other researchers, publishers, and/or edtech developers hoping to apply our techniques to their own math assessment items should consider.

1. GPT will try to give a narrative response explaining the answer it provides. You should set *creativity* to "precise" and *max response size* to "short" in the settings for the GPT Sheets API integration, and you should specify the format in which you want the responses (e.g. what labels to use, how to deliminate terms in a list).
2. You must instruct GPT to not solve the assessment item in your prompt. It will naturally want to consider the answer to the question in its response.
3. You will likely need to go through a manual cleaning process to remove predictable errors such as inconsistent use of periods or elaborations on/justifications for the response.
4. If a single prompt does not produce satisfactory results, a two-step prompt may provide a solution. In the case of context, we were able to distinguish School Math and Symbolic math with a second, unrelated prompt focusing on different problem features.

There are a number of creative ways in which LLMs can augment or improve educational research and impact student learning outcomes. Applications such as the one described in this paper represent a fruitful direction for such exploration with minimal risk.

5. Acknowledgements

The research reported here was supported by philanthropic foundations.

6. References

[1] Huang, Z., Liu, Q., Chen, Y., Wu, L., Xiao, K., Chen, E., Ma, H., & Hu, G. (2020). Learning or Forgetting? A Dynamic Approach for Tracking the Knowledge Proficiency of Students. *ACM Transactions on Information Systems*, *38*(2), 1–33. https://doi.org/10.1145/3379507

[2] Boonen, A. J. H., de Koning, B. B., Jolles, J., & van der Schoot, M. (2016). Word Problem Solving in Contemporary Math Education: A Plea for Reading Comprehension Skills Training. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00191

[3] Hegarty, M., Mayer, R. E., & Green, C. E. (n.d.). *Comprehension of Arithmetic Word Problems: Evidence From Students' Eye Fixations*. 9.

[4] Lin, X., Peng, P., & Zeng, J. (2021). Understanding the relation between mathematics vocabulary and mathematics performance: A meta-analysis. *The Elementary School Journal*, *121*(3), 504-540

[5] Hughes, E. M., Powell, S. R., & Lee, J.-Y. (2020). Development and Psychometric Report of a Middle-School Mathematics Vocabulary Measure. *Assessment for Effective Intervention*, *45*(3), 226–234. https://doi.org/10.1177/1534508418820116

[6] Lewis, AB. (1989). Training students to represent arithmetic word problems. *Journal of Educational Psychology*, *81*(4), 521. https://doi.org/10.1037/0022-0663.81.4.521

[7] Shum, H. Y., & Chan, W. W. L. (2020). Young children's inhibition of keyword heuristic in solving arithmetic word problems. *Human behaviour and brain*.

[8] Boonen, A. J. H., van der Schoot, M., van Wesel, F., de Vries, M. H., & Jolles, J. (2013). What underlies successful word problem solving? A path analysis in sixth grade students. *Contemporary Educational Psychology*, *38*(3), 271–279. https://doi.org/10.1016/j.cedpsych.2013.05.001

[9]Swanson, H. L., & Beebe-Frankenberger, M. (2004). The Relationship Between Working Memory and Mathematical Problem Solving in Children at Risk and Not at Risk for Serious Math Difficulties. *Journal of Educational Psychology*, *96*(3), 471–491. https://doi.org/10.1037/0022-0663.96.3.471

[10] Talarian. (2023). *Use chatgpt in Google Sheets and DocsSupports all models: GPT-4 (if you have access), chatgpt (GPT-3.5-turbo) and GPT-3*. ChatGPT for Google Sheets and Docs. https://gptforwork.com/

[11] Walkington, C., Sherman, M., & Petrosino, A. (2012). "Playing the game" of story problems: Coordinating situation-based reasoning with algebraic representation. *The Journal of Mathematical Behavior*, *31*(2), 174–195. https://doi.org/10.1016/j.jmathb.2011.12.009

[12] Martiniello, M. (2008). Language and the performance of English-language learners in math word problems. Harvard Educational Review, 78(2), 333-368