Citation: Butler, R., Mauntel, M., Aswad, M. Choi, B., & Sharpe, S. (2024). Prompting a Large Language Model for Meta-Synthesis Data Extraction. Paper presented at AERA 2024.

Prompting a Large Language Model for Meta-Synthesis Data Extraction

Current developments in artificial intelligence (AI) tools, such as ChatGPT, have caught the attention of education researchers seeking to expedite time-intensive research processes. In this paper, we explore how one such tool, PDFGear, can be used to extract information from studies to generate memos for a meta-synthesis. We present a prompting protocol for this task and discuss advantages and disadvantages of its implementation. Notably, we advocate for a process that integrates AI tools and researcher expertise in this process. While we encountered challenges to AI memo generation, such as false information and lack of information, this tool vastly decreased the time needed to generate initial memos.

Objectives

Recent advances in artificial intelligence (AI) technologies have produced computer-based tools capable of analyzing and producing textual data in ways that are starkly comparable to human use of language. Since education research is often a time-intensive process involving the analysis of large volumes of language-based data, education researchers have begun to advocate for the use of AI tools to expedite analyses of textual data (Kučak et al., 2018; Longo, 2019; Williamson & Eynon, 2020). While such tools can be advantageous in text analysis, there are also challenges to their use, such as provision of incorrect information and difficulty of cross-context application (Fesler et al., 2019; Fischer et al., 2020). Currently, researchers in the field are seeking methods to mitigate these challenges so that AI tools can be effectively incorporated into data analysis methods.

One methodology apt to reap the potential benefits of AI tools is meta-synthesis because this approach is time-intensive and complex. Meta-synthesis is the process of systematically selecting research material and collectively "synthesizing, analyzing, and interpreting [their] findings" (p. 319, Barry & Thunder, 2016). Accordingly, the data extraction process for a meta-synthesis requires researchers to summarize large amounts of qualitative data, often in the form of structured memos. Our study explores ways to expedite this data extraction process with recent advances in AI. Specifically, we ask, how can an LLM be prompted to extract data for meta-synthesis memos, and what are some of the affordances and limitations of AI memo generation?

In this paper, we describe our use of the AI chatbot embedded within PDFGear to generate memos for a meta-synthesis, comparing these memos to human-generated memos. This comparison will contribute greater insight into the appropriateness of this specific tool for qualitative data analysis as well as a potential means of expediting meta-synthesis data extraction without loss of quality.

Literature and Theoretical Framing

Within the past half-decade, exceptional progress has been made in developing AI tools for text analysis. This application of AI, known as natural language processing (NLP), has been used to respond to human prompts, summarize text, retrieve information, and analyze affective components of text (Chowdhary, 2020). One type of NLP is a Large Language Model (LLM), which uses vast amounts of human-generated text to construct models of human language,

enabling them to coherently produce novel texts (Wei et al., 2022). A notable LLM which has enjoyed a wide public reception is ChatGPT, developed by OpenAI. The latest version of this LLM can respond to a wide array of human requests (e.g., writing emails, generating trip itineraries, outlining essays) in language which is nearly indistinguishable from human-generated texts (Lund & Wang, 2023).

While ChatGPT can produce coherent texts, these texts are not always informationally accurate. This information is especially dubious in tasks requiring logical inferences or context-specific details (Lund & Wang, 2023). Despite this, it is evident that ChatGPT can produce semantic information beyond syntactic relationships, and it has been shown to produce accurate content summaries for various text types, including social media posts, news articles, and verbal dialog (Bradley et al., 2022; Yang et al., 2023).

Alshami et al. (2023) recently used ChatGPT to identify, screen, and extract data from a number of articles for a systematic review. In extracting information, the authors note the importance of "human oversight and critical evaluation" (Alshami et al., 2023, p.37) in interpreting text produced by ChatGPT. We take this informational fallibility of LLMs seriously, operating under an agnostic theoretical perspective that assumes that AI tools must be carefully tailored for the specific task at hand (Chen et al., 2018; Grimmer et al., 2021). Accordingly, in this study, we focus our efforts on exploring how AI tools can be used in conjunction with human discretion to describe qualitative data, rather than suggesting a universal prompting scheme for data extraction.

Methods and Data Sources

The current research is part of a larger research project to analyze and synthesize 40 years of empirical research on algebra teaching approach across multiple algebra areas – early algebra, pre-algebra, and algebra (i.e., grades K-12) and across seismic changes in the mathematics education field (i.e., standards/policies). The larger research project aims to conduct meta-research (systematic review, meta-analysis, and meta-synthesis) to identify what type of algebra teaching approach in the classroom works, for whom, and under what conditions.

Part of the meta-synthesis process is the extraction of relevant qualitative data from the selected articles for further analysis (Ong et al., 2023). For our project, the relevant portions of extraction include the research question from the article, a description of the intervention, summary of the results, and the implications of the study. We selected six papers that have passed an abstract and full-text screening for the larger project to be included in this study. For each study, a team member created a memo that summarized the relevant portions of the paper. Another team member then reviewed this memo to ensure accuracy.

In this paper, we utilize AI tools to extract information from these six articles related to their research purpose and classroom intervention in a way which is consistent with the human memoing process. We chose PDFGear as the interface to generate AI memos since it is a publicly available free tool designed to summarize the text of PDFs using ChatGPT 3.5 (PDF Gear Tech Pte Ltd., 2023). This tool provides page numbers of the PDF to support the claims made in its responses, enabling researchers to check the accuracy of this information. To interface with this tool, we follow Eager and Brunton's (2023) framework for prompt engineering in education. This framework is iterative; the researchers generate initial prompts to communicate outcomes for the AI to produce, then refine these prompts based on the AI's response. In our results below, we describe the prompting protocol developed from this iterative methodological approach, along with examples of the AI memos produced from this protocol.

Results

Our protocol (Figure 1) does not prompt the AI to independently return a completed memo in response to specific prompts, but instead relies upon the content specific expertise of the researcher to filter AI responses into coherent and accurate memos through an iterative process. Researchers pose initial questions, then reflect on these responses with attention to key words and phrases which are repeated across responses. They use their expertise to craft additional prompts using these key words and phrases. Researchers identify sections that contain repeated and accurate information to generate the memo. The resulting memo is the genesis of the interaction between the AI and the expertise of the researcher (Figure 2).

This active role of the researcher in creating AI memos is motivated by several challenges noted by our team in the initial prompting of the AI. Our first challenge was the generation of incorrect information, including statements about participants not included in the study, generation of quotes that do not appear in the paper, and misinformation about the study's focus. While misinformation about specific details was relatively rampant within individual responses, we noticed that, across responses, words and phrases were consistent descriptions emerged. Further, when asked about these repeated phrases, the AI returned reliable information.

In addition to incorrect information, we also struggled with a lack of essential information in responses. For example, when we asked each of the intervention questions from our protocol about Bulgar's (2003) paper, only broad descriptions of the classroom intervention were given, such as "the study implemented instructional strategies related to creating problem-based activities and utilizing manipulatives to promote conceptual understanding of mathematical ideas related to fractions." While this is an accurate summary, it lacks details of the intervention. However, these details were correctly provided when the researcher asked the AI to describe the classroom activities, with the AI responding

> "The task that students did involved solving a problem called 'Holiday Bows' in order to elicit ideas relating to division of fractions. In this problem, students were given varied lengths of ribbon and were instructed to make bows of different fractional sizes from each length of ribbon. Actual ribbons, precut to the indicated sizes, were available to the students".

The initial description was vague, but with the researcher's attention to the use of an activity in the study, these details were obtained. Attention to key words and phrases, as identified by an expert in education research, is essential in our prompting protocol as they help direct the AI toward the necessary information.

The lack of essential information in responses was also addressed through the design of initial prompts. This is evident in our prompting for the intervention, as we used our expertise as researchers to specify what we mean by "intervention" through a series of questions about different intervention types. Again, it was essential for us to use our knowledge as mathematics education researchers to obtain the information required for the memos.

Despite these challenges, our prompting protocol helped us generate memos that were similar to human-generated memos. Figure 3 provides several side-by-side comparisons of AI-generated memos and human-generated memos. Notably, the AI showed inconsistency with identifying explicitly stated research questions, as seen for the Ventura et al. (2021) paper, and occasionally generated research questions that were not stated by the authors but provided accurate information about the paper, such as seen for the Paoletti et al. (2019) paper. While the AI descriptions were typically longer than the human-generated summaries, these memos shared

key words and ideas. For example, in the Moss and Lamberg (2019) intervention memos, both the human and AI memos contain “Kaput's framework of algebraic thinking,” “Common Core State Standards for Mathematics,” and “expressions and equations” as essential components of the intervention.

While we encountered difficulties in our initial use of AI for qualitative memo writing, the use of our protocol has streamlined our memoing process. When generated by a human alone, memos took three to four hours to write, not including time for verification. Using our protocol, memos took approximately one hour to draft, including time used to verify the information provided in these memos. Checking that the AI-supplied information was consistent with the given paper was an easy task since PDFGear references specific pages in its responses. Researchers were able to scan these referenced pages and use a document search tool to verify information in the AI memos quickly.

Significance

Our prompting protocol and AI-generated memos show the utility of AI tools in the qualitative memoing process. While this technology is not at the point of independently generating summaries of qualitative studies, it is a powerful tool when used in conjunction with the discernment of a skilled researcher. Using our expertise as mathematics education researchers, we were able to effectively communicate complex constructs from the field to the AI, direct the AI toward relevant information within qualitative papers, and critically assess the AI-generated responses for accuracy and significance. While we note similar difficulties in information extraction as Alshami et al. (2023), such as reliability and absence of information, our initial results provide an important example of how human expertise, specifically within the context of education research, can be used in conjunction with AI tools to expedite the research process.

In our continued application of AI to qualitative memo writing, we will expand our prompting protocol to extract information related to the theoretical framing and results of the study. Just as we tailored a specific set of prompts to communicate what we as researchers meant by the word “intervention”, specific prompts will need to be developed to communicate what is meant by “theoretical framing” and “results.” We are especially interested in exploring how to extract mathematical symbols and information contained in figures from studies using AI. Through continued prompt development, we aim to further contextualize the nature of a productive relationship between researchers and AI tools to increase efficiency and quality within meta-synthesis approaches.

Acknowledgments

This research was supported by the National Science Foundation DRK-12 grant #2142659. Any opinions, findings, or conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References (*studies were memoed)

Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A. E., & Zayed, T. (2023). Harnessing the Power of ChatGPT for Automating Systematic Review Process: Methodology, Case Study, Limitations, and Future Directions. *Systems*, *11*(7), 351. https://doi.org/10.3390/systems11070351

Berry, R., & Thunder, K. (2013). The promise of qualitative metasynthesis: Mathematics experiences of black learners. *Journal of Mathematics Education at Teachers College*, *3*(2), 43-55. https://doi.org/10.7916/jmetc.v3i2.757

Bradley, T. D., Terilla, J., & Vlassopoulos, Y. (2022). An enriched category theory of language: from syntax to semantics. *La Matematica*, *1*(2), 551-580. https://doi.org/10.1007/s44007-022-00021-2

*Bulgar, S. (2003). Children's sense-making of division of fractions. *The Journal of Mathematical Behavior*, *22*(3), 319-334. https://doi.org/10.1016/S0732-3123(03)00024-5

*Carraher, D. W., Schliemann, A. D., Brizuela, B. M., & Earnest, D. (2006). Arithmetic and algebra in early mathematics education. *Journal for Research in Mathematics education*, *37*(2), 87-115. https://doi.org/10.2307/30034843

Chen, N. C., Drouhard, M., Kocielnik, R., Suh, J., & Aragon, C. R. (2018). Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *8*(2), 1-20. https://doi.org/10.1145/3185515

Chowdhary, K. (2020). Natural language processing. In K. R. Chowdhary (Eds.), *Fundamentals of artificial intelligence* (pp. 603-649). Springer.

Eager, B., & Brunton, R. (2023). Prompting higher education towards AI-augmented teaching and learning practice. *Journal of University Teaching & Learning Practice*, *20*(5), 02. https://doi.org/10.53761/1.20.5.02

Fesler, L., Dee, T., Baker, R., & Evans, B. (2019). Text as data methods for education research. *Journal of Research on Educational Effectiveness*, *12*(4), 707-727. https://doi.org/10.1080/19345747.2019.1634168

Fischer, C., Pardos, Z. A., Baker, R. S., Williams, J. J., Smyth, P., Yu, R., Slater, S., Baker, R., & Warschauer, M. (2020). Mining big data in education: Affordances and challenges. *Review of Research in Education*, *44*(1), 130-160. https://doi.org/10.3102/0091732X20903304

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, *24*, 395-419. https://doi.org/10.1146/annurev-polisci-053119-015921

Kučak, D., Juričić, V., & Đambić, G. (2018). Machine Learning in Education - A Survey of Current Research Trends. In B. Katalinic (Ed.), *Proceedings of the 29th DAAAM International Symposium* (pp.0406-0410). DAAAM International, Vienna, Austria. https://doi.org/10.2507/29th.daaam.proceedings.059

Longo, L. (2019). Empowering qualitative research methods in education with artificial intelligence. In A. P. Costa, L. P. Reis, & A. Moreira (Eds.), *World Conference on Qualitative Research* (pp. 1-21). Cham: Springer International Publishing.

Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: how may AI and GPT impact academia and libraries?. *Library Hi Tech News*, *40*(3), 26-29. https://doi.org/10.1108/LHTN-01-2023-0009

*Moss, D. L., & Lamberg, T. (2019). Conceptions of expressions and equations in early algebra: A learning trajectory. *International Journal for Mathematics Teaching and Learning*, *20*(2), 170-192.

Ong, M., Jaumot-Pascual, N., Torres-Gerald, L., Martínez-Gudapakkam, A., & Silva, C. B. (2022). Institute for meta-synthesis user guide: Eight applied modules to learn and practice qualitative meta-synthesis. TERC.

*Paoletti, T., Vishnubhotla, M., & Mohamed, M. (2019). Inequalities and Systems of Relationships: Reasoning Covariationally to Develop Productive Meanings. In S. Otten, A. G. Candela, Z. de Araugo, C. Haines, & C. Munter (Eds.), *Proceedings of the Forty-First Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education* (pp.157-166). North American Chapter of the International Group for the Psychology of Mathematics Education.

PDF Gear Tech Pte Ltd. (2023). PDFGear (July '23 version). https://www.pdfgear.com/.

Thunder, K., & Berry, R. Q. (2016). The promise of qualitative metasynthesis for mathematics education. *Journal for Research in Mathematics Education, 47*(4), 318-337. https://doi.org/10.5951/jresematheduc.47.4.0318

*Tondorf, A., & Prediger, S. (2022). Connecting characterizations of equivalence of expressions: design research in Grade 5 by bridging graphical and symbolic representations. *Educational Studies in Mathematics*, *111*(3), 399-422. https://doi.org/10.1007/s10649-022-10158-0

*Ventura, A. C., Brizuela, B. M., Blanton, M., Sawrey, K., Gardiner, A. M., & Newman-Owens, A. (2021). A learning trajectory in kindergarten and first grade students' thinking of variable and use of variable notation to represent indeterminate quantities. *The Journal of Mathematical Behavior*, *62*, 100866. https://doi.org/10.1016/j.jmathb.2021.100866

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zohu, D., Metzler, D., Chi, E.H., Hashimoto, T., Hashimoto, O., Liang, P., Dean, J., Fedus, W. (2022). Emergent abilities of large language models. arXiv. https://doi.org/10.48550/arXiv.2206.07682

Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learning, Media and Technology*, *45*(3), 223-235. https://doi.org/10.1080/17439884.2020.1798995

Yang, X., Li, Y., Zhang, X., Chen, H., & Cheng, W. (2023). Exploring the limits of chatgpt for query or aspect-based text summarization. arXiv. https://doi.org/10.48550/arXiv.2302.08081

Figures and Tables

Figure 1: Prompting Protocol

Prompting For Interventions and Results

1) Prompting for purpose
   a) Examine the auto-generated summary for key words, ask the auto suggested questions
   b) Ask the following questions
      i) What is the purpose of the study?
      ii) What are the study's research questions?
   c) Reflect on the responses to questions asked in 1a and 1b, ask for more detail about key words if needed. Stop this process when new prompts fail to generate new information.
2) Prompting for the intervention
   a) As the following questions (one at a time) to determine the intervention type
      i) Does the study implement instructional strategies?
      ii) Does the study implement learning strategies?
      iii) Does the study implement a curriculum?
      iv) Does the study implement technology?
      v) Does the study implement tutoring?
      vi) Does the study implement manipulatives?
      vii) Does the study implement teacher development?
   b) Reflect on the responses to each of these prompts, looking for keywords used across prompts and facts which are consistently provided across prompts
   c) Once the intervention has been determined, ask for a specific description based on keywords used in previous responses (e.g., "describe the task in the study", "describe the applet used in the study"). Stop this process when prompts fail to generate new information.

Figure 2: AI Prompting Process

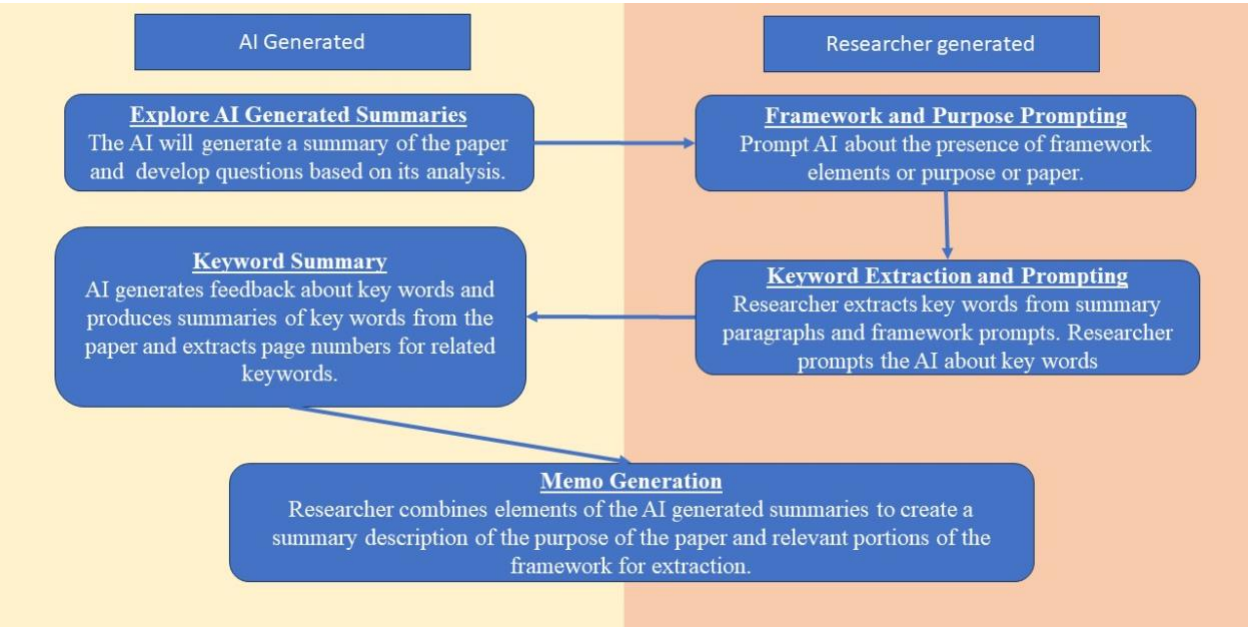

Figure 3: AI and Human Generated Memos

Paper – Memo Component	AI-Generated Memo	Human-Generated Memo
Moss & Lamberg (2019) - Intervention	Yes, the study implemented instructional strategies during the teaching experiment. The experiment involved a whole-class teaching approach using design research. The instructional strategies were based on Kaput's framework of algebraic thinking and the objectives of the sixth-grade Common Core State Standards for Mathematics. During the experiment, decisions were made about the teaching approaches, the types of activities and tasks, and the order and ways that would address the objectives of the research. The study involved a teach/reflect/plan approach, making design decisions based on student thinking and responses. The researchers recorded their design decisions, the teaching approaches used in each session, and the types of activities and tasks performed during lessons. The data collected during the experiment, including video of the teaching sessions, video of student discussions and comments, and student work, were used to analyze the effectiveness of the instructional strategies The study aimed to observe how sixth-grade students came to understand expressions and equations through a whole-class teaching experiment using design research. The experiment focused on developing an instructional sequence based on Kaput's framework of algebraic thinking and identifying the types of student thinking that emerged through tasks that were presented. The activities and tasks used during the experiment were designed to target these different types of thinking. The study used a variety of approaches, including problem-solving activities and tasks, group discussions, peer interactions, feedback, and reflection. The data collected during the experiment included video of the teaching sessions, video of the student discussions and comments, and video and field notes of the teacher/researcher debriefing interviews, lesson plans, and student work. The qualitative data generated by this study was voluminous, and the retrospective analysis consisted formal analysis of the video recordings, student work, and field notes.	A hypothetical learning trajectory for sixth grade algebra was developed by integrating research on how students learn algebra and the Common Core State Standards for Mathematics in the domain of Expressions and Equations. The trajectory was based on Kaput's (1999) framework for algebraic thinking, which emphasized the connections between arithmetic and algebra, the meaning furl use of symbols, and mathematical modeling.

Paoletti et al. (2019)- Purpose	The authors were trying to answer several research questions related to students' covariational reasoning and how it is affected by systems of relationships and inequalities. The questions include: 1) How can systems of relationships and inequalities be used to develop students' covariational reasoning? 2) How do students reason about systems of relationships and inequalities involving two quantities with respect to a third quantity? 3) How do students' constructions and use of representations (pictorial, graphic, symbolic, and verbal) support their reasoning about systems of relationships and inequalities involving two quantities? 4) How can teachers facilitate students' reasoning about systems of relationships and inequalities involving two quantities using physical manipulatives? 5) What kinds of covariational reasoning do middle-school students display when solving tasks involving systems of relationships and inequalities? Therefore, the authors were trying to investigate various aspects of covariational reasoning and how it is affected by different factors such as representations, systems of relationships and inequalities, and the use of physical manipulatives.	The research problem, purpose, questions, and hypothesis are not mentioned directly, but in this article, the authors mentioned that "we extend Olive and Caglayan's (2008) work by examining ways to support middle-grades students in reasoning quantitatively and covariationally to construct, graph, and determine solutions to systems of relationships."
Ventura at al. (2021)- Purpose	The purpose of the study described in this article is to investigate Kindergarten and first grade students' thinking about problems that can be symbolized with the expression x + b, where x is an indeterminate. The study aims to validate and refine the already existing empirically-developed learning trajectories for students' algebraic thinking. The learning trajectory used in this study for the development of algebraic thinking for young students was developed by Blanton, et al. (2017). The authors of this study validated and refined the existing learning trajectory developed by Blanton, et al. (2017) by testing and extending it with teaching experiments.	RQ1. In what ways do Kindergarten and first grade students progress in their understanding of variable and use variable notation as a way to represent indeterminate quantities and relationships between indeterminate quantities in algebraic expressions? RQ2. What are the connections between an existing learning progression describing first grade children's thinking about variable and variable notation in functional relationships and the ways in which Kindergarten and first grade children understand variable and variable notation in algebraic expressions? p.2