Large Language Models (LLMs) for Evidence Synthesis: An Exploratory Evaluation and A New Approach for Automated Data Extraction

Yuchen Zhang[1], Nanyu Luo[1], Hajung Kim[1], Linxin Li[2], Linfeng Gao[3],

Jiayi Han[4], Shi'ting Chen[4], Xiaoya Zhang[5], Jinbo He[6,*], Feng Ji [1,*]

[1] Department of Applied Psychology and Human Development, University of Toronto, Toronto, Canada

[2] Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, Canada

[3] Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada

[4] Division of Applied Psychology, School of Humanities and Social Science, The Chinese University of Hong Kong, Shenzhen, Guangdong, 518172, China

[5] Department of Family, Youth and Community Sciences, University of Florida, USA

[6] Department of Biosciences and Bioinformatics, School of Science, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215123, China

[*]Correspondence concerning this article should be addressed to Feng Ji, PhD, Department of Applied Psychology and Human Development, University of Toronto, Toronto, Canada; E-mail: f.ji@utoronto.ca. And Jinbo He, PhD, Department of Biosciences and Bioinformatics, School of Science, Xi'an Jiaotong-Liverpool University, Suzhou, Jiangsu, 215123, China; Email: Jinbo.He@xjtlu.edu.cn; anlfhe@gmail.com

Declaration of Conflicting Interests

The author(s) report having no conflicts of interest with respect to contents, authorship, or publication of this article.

Funding

This project was supported by the Connaught Fund (Grant Number 520245) and Social Sciences and Humanities Research Council (SSHRC) of Canada (Grant Number 215119, 00169), Seed Grant of the International Network of Educational Institutes (Grant Number 522011).

Abstract

Large language models (LLMs) are increasingly used in scientific research for their strong general problem-solving capabilities. Data extraction remains one of the most time- and labor-consuming steps in evidence synthesis (ES), making LLMs a promising tool with improved efficiency and accuracy. Our study evaluates the performance of different LLMs and proposes a novel method, **Divide, Conquer, then Recheck** (**DCR**), to optimize for LLM-based data extraction in ES. Multiple LLM foundational models were compared through accuracy, precision, recall, and F1-score. We find that GPT-4o demonstrates notably better performance across most variables compared to ChatPDF, Bing Chat, and GPT-4. The proposed DCR method powered by GPT4-o achieved higher accuracy in most structured data extraction and the few-shot prompting strategy further improved performance on complex information (e.g., correlation coefficient). These findings highlight the potential of using LLMs in ES research.

Keywords: Generative AI, Large Language Models, Responsible AI, ChatGPT, Evidence Synthesis, Meta-analysis, Research Methods

Large Language Models (LLMs) for Evidence Synthesis: An Exploratory Evaluation for Automated Data Extraction and A Novel Method with Improved Performance

1. Introduction

Evidence synthesis (ES) is a systematic research process of integrating information from multiple studies to provide a comprehensive understanding of a specific topic. It involves methodical and reproducible approaches to identify, evaluate, and combine evidence on a given research question from diverse sources (Ades & Sutton, 2006). Beyond its methodological rigor, ES also provides a practical and structured approach to evaluate the effectiveness and risks of interventions and treatments, guiding research, policy, and decision-making (Briner et al., 2012; Khan et al., 2003; Ohlsson, 1994), and has become increasingly influential across the social, behavioral, and medical sciences (Littell et al., 2008). Despite their significance, ES is often time-consuming and resource-intensive, resulting in delays that may render findings outdated by the time they are published (Blaizot et al., 2022; Edwards et al., 2023; Hill et al., 2023). In this study, we focus on whether, what, and how recent advances of large language models (LLMs) can address these limitations and improve ES.

1.1. Introduction to Evidence Synthesis

Evidence synthesis (ES), such as systematic reviews and meta-analyses, typically involves several key steps: formulating the review question, developing a protocol, identifying and selecting relevant studies, critically appraising their quality, analyzing and synthesizing the findings, and disseminating the review results (Briner et al., 2012). Among these, a crucial step is data extraction, the process of retrieving and organizing key information from primary studies. This step functions as a bridge between study selection and critical appraisal, providing the foundational data necessary to evaluate the quality, reliability, and relevance of the included studies.

However, data extraction and coding is also one of the most time-consuming and resource-intensive stages of the ES process. This step is typically conducted manually, making it both labor-intensive and prone to human error. Studies have shown that data extraction errors are common, with rates reported as high as 63% depending on the complexity of the data (Mathes et al., 2017). These errors may stem from misclassification of statistical values, misinterpretation of findings, or simple data entry mistakes. In addition to undermining the accuracy of ES, manual data extraction contributes to delays in result dissemination and reduces reproducibility, raising concerns about the efficiency and reliability of ES.

1.2. Current Research and Limitations on LLMs in ES

Recent advances in artificial intelligence (AI) have introduced large language models (LLMs) as a promising tool to automate data extraction in ES. These models can process natural language, recognize complex patterns, and perform extraction tasks with minimal training (OpenAI, 2023; Bubeck et al., 2023; De Angelis et al., 2023; Paruchuri et al., 2024). A growing body of studies has started to explore the potential utility of LLMs in academic research settings (e.g., Mammides & Papadopoulos, 2024), including forecasting and anomaly detection (e.g., Su et al., 2024), sentiment analysis (e.g., Susnjak, 2024), time series analysis (e.g., Zhang et al., 2024), power analysis (e.g., Kim et al., 2024), as well as, ES (e.g., Hill et al., 2024; Polak et al., 2024; Wang & Luo, 2024).

Recent research on LLMs in ES has primarily focused on automating tasks such as screening publications against inclusion/exclusion criteria (e.g., Wang et al., 2024) and data

extraction (Polak & Mogan, 2024; Wang & Luo, 2024). These approaches have shown particularly strong performance in healthcare and materials science, especially when extracting structured and well-formatted data (e.g., binary variables, material properties; Ekuma, 2024; Wang & Luo, 2024). However, the accuracy rate tends to decline when dealing with continuous or complex data types (e.g., Sun et al., 2024). Notably, the majority of these studies have focused on highly structured research designs, such as randomized controlled trials (RCTs), where data formats are relatively more standardized and thus more amenable for automation.

In contrast, the application of LLMs in social and behavioral science remains largely underexplored. Social and behavioral research often involves complex constructs, diverse methodological designs, and context-dependent reporting (Sanbonmatsu et al., 2021), making it inherently difficult to extract and standardize data. Since LLMs are pretrained on broad, general-purpose datasets, they may lack the domain-specific precision required to accurately handle highly specialized content (Susnjak, 2024). As a result, automated extraction and coding face particular challenges when dealing with embedded statistics, complex quantitative indicators, or unstructured qualitative data.

Moreover, while prior studies have examined the feasibility of LLM-based data extraction and some have compared extracting performance of different LLMs (Dagdelen et al., 2024; Celikten & Onan, 2025), relatively few have systematically assessed data extraction performance under schema constraints, including mapping diverse data types into structured coding schemes. Existing evaluations are often limited by small sample sizes (e.g., validation of results extracted from 15 RCT medical papers, Vidal Perez, 2024), further constraining the generalizability of findings across broader ES contexts. To address these gaps, our study explores the capabilities and methodological innovations of using LLMs in extracting and coding diverse data formats and types within the social and behavioral sciences.

Hallucination and lack of accuracy also remain critical challenges in LLM-based data extraction and coding, particularly within the context of ES. To date, although LLMs have offered substantial potential to streamline information retrieval, they are still susceptible to generating inaccurate or fabricated outputs, especially when dealing with very large context windows (Liu et al., 2024) and multi-modal data when reviewing articles. These limitations often result in missing or inconsistent extractions that compromise the reliability of ES. One contributing factor may be the sequential nature of LLM text generation, compounded by API constraints such as Tokens Per Minute (TPM), which may ultimately produce incorrect outputs. Yet, few studies have examined how to systematically identify and resolve such discrepancies. As a result, there is a growing need for approaches that not only automate data extraction and coding to improve efficiency but also incorporate robust validation methods to ensure the accuracy and reliability of extracted outputs.

1.3. The Objective of the Study

Building on a standardized coding scheme in social behavioral research, we focus on the topic of body image and selected 60 papers and evaluate the performance of different LLMs in extracting and coding the information from these papers, and we compared different models (i.e., GPT-3.5, GPT-4, GPT-4o) based on precision, accuracy, recall, and F1-score. Meanwhile, building on the comparative evaluation, we then propose an improved automated method, **Divide, Conquer, then Recheck** (**DCR**), to enhance the extraction performance and reduce reliance on human validation.

Based on our research purpose, we aim to answer the following research questions: (1) To what extent can different LLMs autonomously and accurately extract and code social and behavioral data from empirical studies? (2) How does the extraction and coding precision improve using our proposed method? (3) Could our proposed method effectively validate and recheck outputs?

2. Method

2.1. Study Design

In order to evaluate the performance of different LLMs in data extraction and coding during the ES process, we build upon an existing well-cited meta-analysis study (He et al., 2020) and use LLMs to extract and code information from the paper identified by this meta-analysis. This paper involves 60 papers on the topic of body image, specifically including body appreciation and BMI. We chose this paper as a gold standard due to its relevance to the fields of social and behavioral sciences and its high quality, as the original coding was conducted by multiple researchers and underwent rigorous peer review. Additionally, it contains a diverse set of qualitative and quantitative variables as presented in Table 1.

Table 1: Summary of Extracted and Coding Information.

Data Types	Extracted Variables (Coding Schemes)
Qualitative (N = 7)	*author(s);* type of publication (*ptype*; journal or dissertation); country (*region*; North America, Europe, South America, or Asia); survey method (*smethod*; online or paper-pencil); source of sample (*source*; primary/middle/high school, college, or community); assessment of body appreciation (*measureba*; BAS or BAS-2); assessment of BMI (*measurebmi*; self-reported or measured);
Quantitative (N = 6)	percentage of participants with a college education or higher (*p2*); sample sizes of participants (*ssize*); percentage of participants identifying as white (*pwhite*); mean age of participants (*mage)*; mean BMI of participants (*mbmi*); and the Pearson correlation between Body Appreciation (BA) and BMI (*r*).

2.1.1 Study 1: Comparison of LLMs for Manual Chatbot-Based Data Extraction

We conducted an evaluation study using ChatPDF (powered by GPT-3.5), Bing Chat (powered by GPT-4), ChatGPT-4, and ChatGPT-4o to evaluate and compare the performance (i.e., accuracy) of different LLMs in the automated data extraction and coding process. Across all phases in Study 1, each article was uploaded individually and manually queried through *chatbots* (*see* Appendix A) to extract and code data. Two independent raters evaluated the accuracy of extracted outputs against a human-coded gold standard. ChatPDF represents an example of the document-centric interface that routes queries to OpenAI models for PDF interrogation, and prior studies show these GPT-4-family systems have already shown potential contributions in ES tasks (Polak & Mogan, 2024; Wang & Luo, 2024), so the results establish a baseline for LLM-

based extraction in social and behavioral ES research and inform further methodological refinement in Study 2.

2.1.2 Study 2: Development of an Automated Extraction and Validation Method

To reduce potential hallucination and further improve factual accuracy, we propose an automated method, the **Divide, Conquer, then Recheck** (**DCR**; *see* Figure 1), which better utilizes the API token limitation and more comprehensively reveals the advanced power of GPT-4o, the model achieving the highest accuracy in Study 1. The DCR method is designed to improve the precision, reliability, and traceability of automated data extraction via LLMs. DCR performs three core functions: (1) dividing full-text articles into smaller and standardized chunks, (2) applying structured and iterative prompting for targeted data extraction, and (3) implementing a systematic rechecking and validating process to identify and resolve discrepancies.

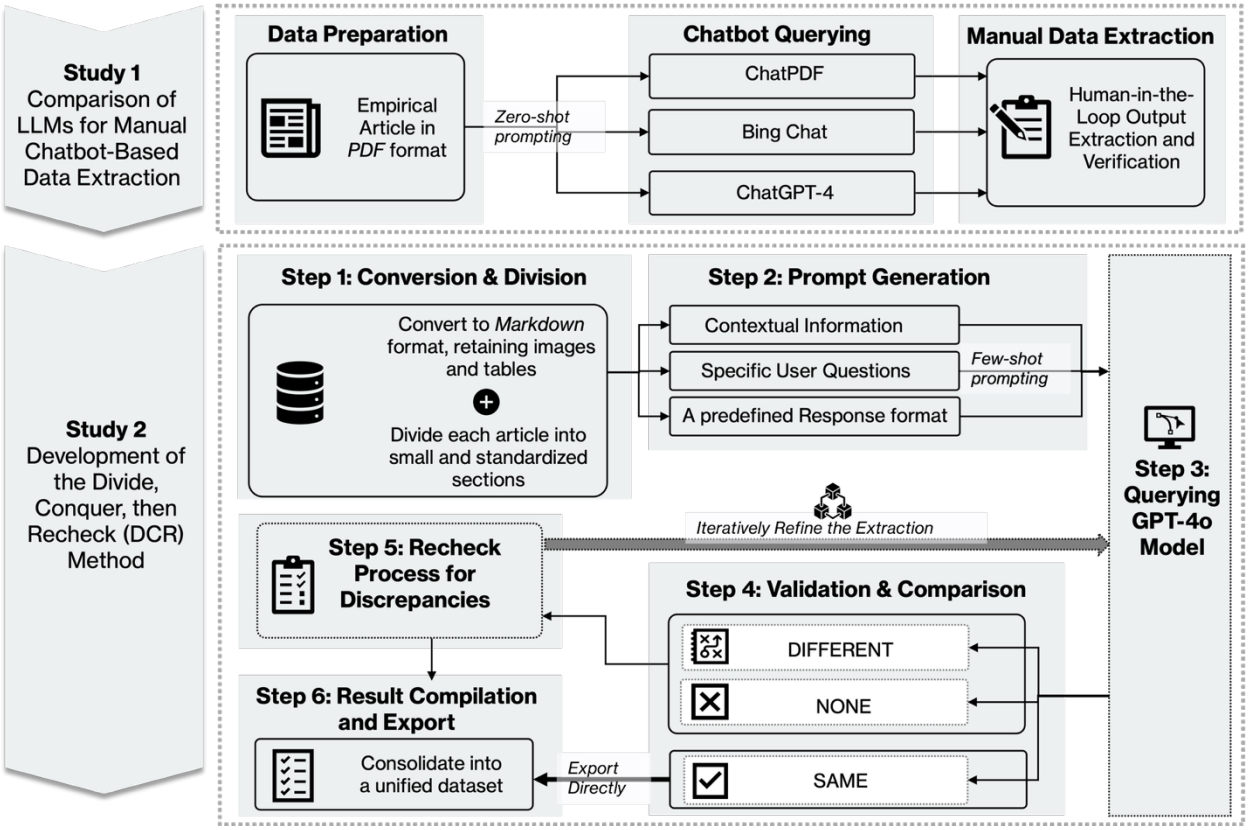

Figure 1: Method Comparison Between Study 1 and Study 2.

Divide and Conquer. To mitigate the token limitation issue and leverage GPT-4o's ability to process multi-modal data, DCR first converts PDF articles into Markdown but preserves images, which reduces file size and storage costs. Articles are then segmented into standardized sections (e.g., introduction, methods), which reduces the input context. Each variable is extracted, via *OpenAI's GPT API*, using strategically tailored prompts, with clear definitions and requirements (e.g., extracted *by sex*) to minimize ambiguity, which might cause hallucination. For complex variables such as correlation coefficients, we implement *few-shot prompting* based on prior low-performing cases (*see* Figure 2, Appendix B).

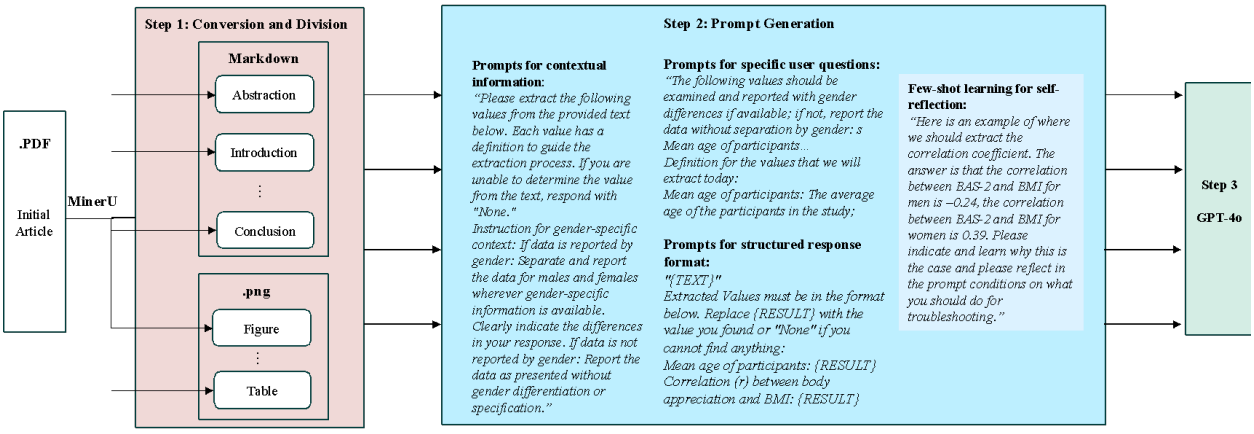

Figure 2: Diagram of Divide, Conquer, and Prompt.

Rechecking. To improve accuracy and consistency, DCR includes a *systematic validation process* that cross-checks extracted values across article sections and categorizes them as *SAME* (aligned), *DIFFERENT* (inconsistent data), or *NONE* (missing data). Cases flagged as *DIFFERENT* or *NONE* trigger a rechecking step, prompting the model to revisit relevant content to clarify discrepancies or locate missing data. This iterative process significantly reduces discrepancy errors, increases coverage, and ensures the robustness and precision of the final results (*see* Figure 3).

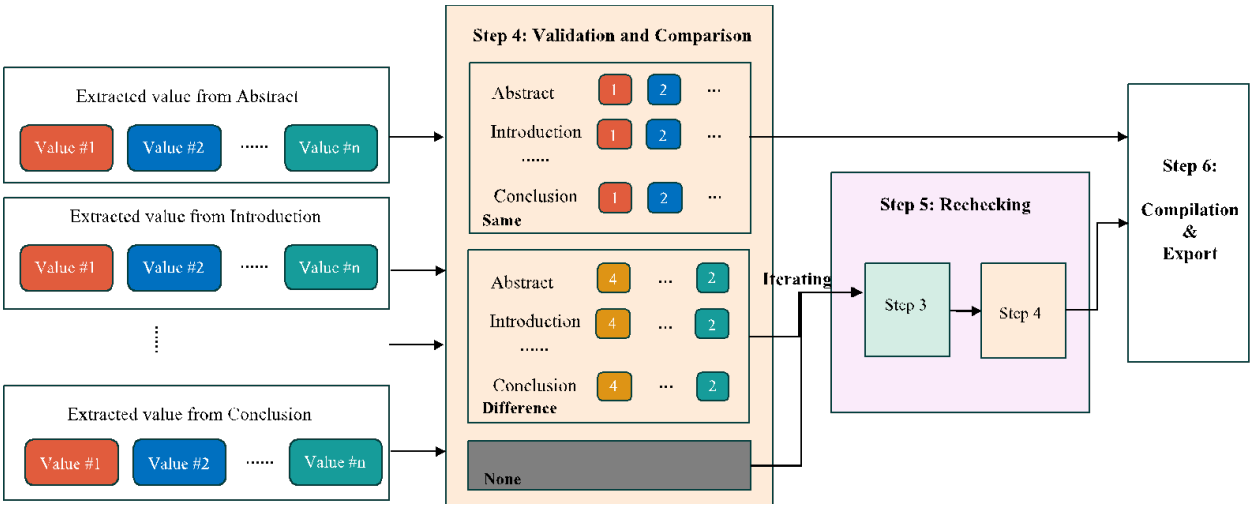

Figure 3: Diagram of Systematic Validation and Rechecking Process.

2.2. Transparency

The Python script for the present study is publicly available on OSF (https://osf.io/zfx9v/?view_only=c8b1b5056fd84093bcf2a343cdfa2d51).

2.3. Key Evaluation Metrics

In Study 1, we primarily use *accuracy* to evaluate the performance of LLMs, which is defined as the proportion of agreement between the LLM-generated outputs and the established gold standard. For each of the 60 selected studies, we calculate the average accuracy across two independent raters, assessing the degree to which ChatPDF, Bing Chat, ChatGPT-4, and GPT-4o responses aligned with the gold standard. This approach allows us to estimate the overall accuracy range and performance for each LLM across all extracted items.

In Study 2, to more comprehensively evaluate the performance of our proposed DCR method, we adopt three more evaluation metrics, including *precision*, *recall*, and *F1-score*. Unlike prior evaluation studies with single metrics (e.g., Wang & Luo, 2024), this approach enables systematic identification of discrepancy issues. To be specific, the *precision* and *recall* scores interpret inconsistent data and missing data, respectively, to measure correctness and completeness, while the *F1-score* summarizes the trade-off between precision and recall, thereby jointly assessing the accuracy, completeness, and reliability of LLM-based extraction compared to the gold standard.

We define three core components: *Correctly Extracted Data* (CED), *Incorrectly Extracted Data* (IED), and *Missing Data* (MD). CED refers to data accurately extracted by the LLM that matches the ground truth (i.e., gold standard results). IED means extracted data that do not match

the ground truth, while MD represents the data present in the ground truth but omitted by the LLM. *Precision* is calculated as the proportion of correctly extracted data (CED) among all extracted data; *Recall* is computed as the proportion of correctly extracted data among all data present in the ground truth; *F1-score* is computed as the harmonic mean of precision and recall. The formulas are as below.

$$Precision = \frac{CED}{CED + IED}\ ;\ Recall = \frac{CED}{CED + MD}\ ;\ F1\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall}$$

3. Results

3.1. Performance of Different LLMs in Data Extraction and Coding

Overall, the accuracy of data extraction and coding varied substantially in Study 1, where we assessed through standard chatbot interfaces. We find that GPT-4o (*total average accuracy* = .80) demonstrate notably better performance across most variables compared to ChatPDF (*total average accuracy* = .54), Bing Chat (*total average accuracy* = .73) and GPT-4 (*total average accuracy* = .79). According to Figure 4, notable improvements were observed in variables such as survey method, source of sample, assessment of BMI, percentage of participants with a college education or higher, and the Pearson correlation between BA and BMI. These results highlight GPT-4o's relative strength in accurately extracting and coding data in ES, which reflects the

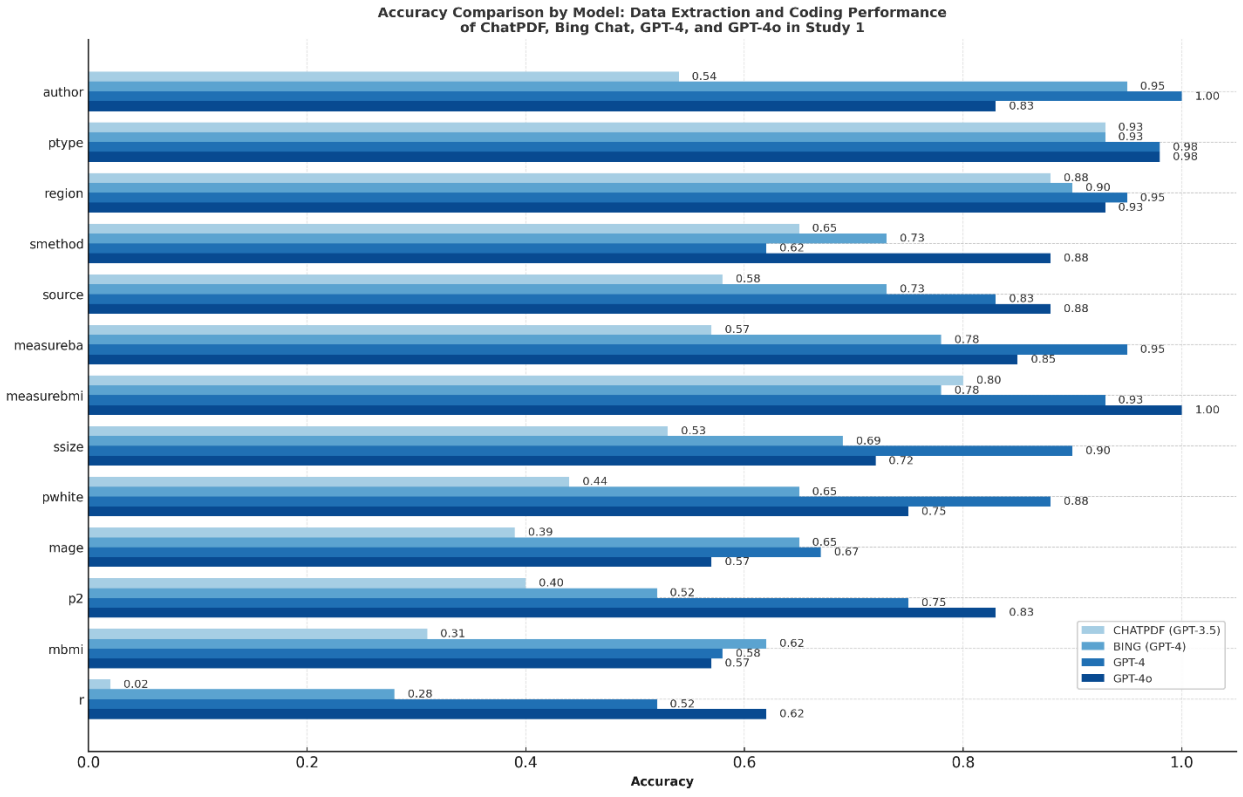

Figure 4: Accuracy Comparison by Models: Data Extraction and Coding Performance of ChatPDF, Bing Chat, GPT-4 and GPT-4o in Study 1.

ongoing evolution of LLM capabilities.

However, we also observed visible differences in the performance between qualitative and quantitative data in Study 1. Generally, qualitative variables were extracted with relatively higher

accuracy compared to quantitative data, which consistently showed lower accuracy due to their complexity (e.g., mean age, mean BMI, and correlation coefficient; *see* Figure 4). Thus, this finding motivated us to proceed with the investigation in Study 2, in which we compared the GPT-4o chatbot with our proposed DCR method to improve extraction accuracy.

3.2. Comparison of the ChatGPT-4o and the DCR Method

Compared to the best-performing model in Study 1 (powered by GPT-4o), the DCR method in Study 2 shows noticeable improvements, with *total accuracy* increasing from .80 to .95 and *total average F1-score* increasing from .89 to .97 (*see* Figure 5). Notably, the total average F1-scores for qualitative data and quantitative data of our extracted variables increase to .97 and .98, respectively, indicating substantial improvements in both data types. These results suggest that our DCR method in Study 2 not only outperforms the best-performing model in Study 1 overall, but also demonstrates greater reliability and consistency across different data types. Specifically,

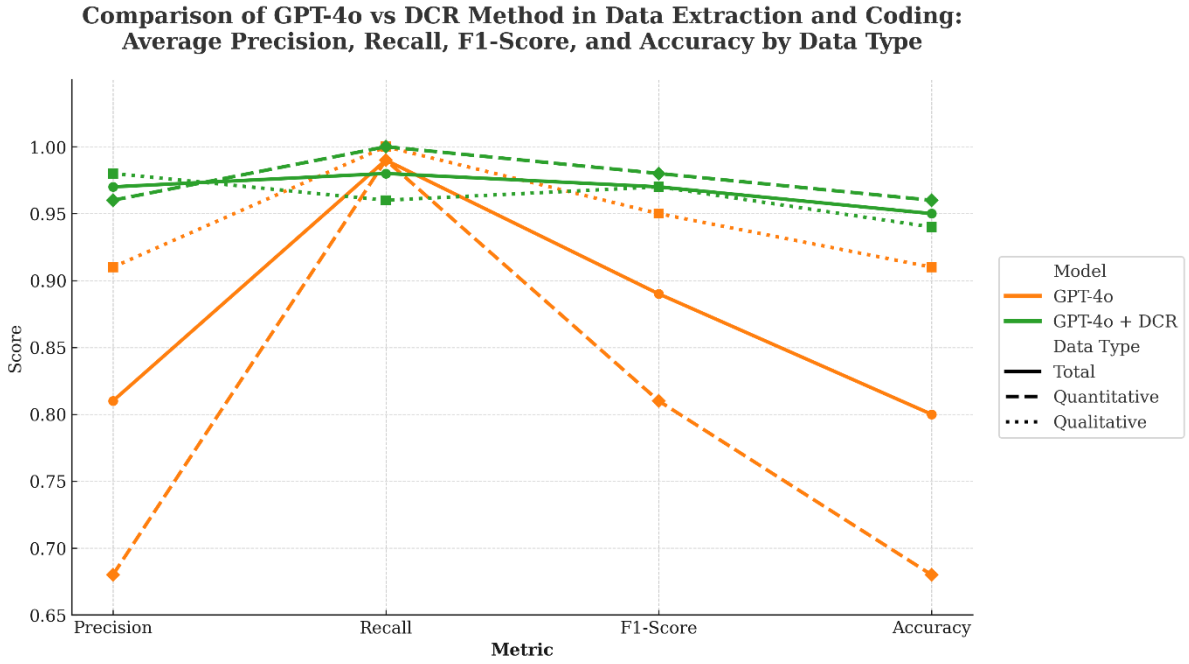

Figure 5: Comparison of GPT-4o vs DCR Method in Data Extraction and Coding Performance: Total Average Precision, Recall, and F1-Score by Data Type.

it helps narrow the accuracy gap in overall quantitative data extraction and coding.

As shown in Figure 5, the DCR method in Study 2's total average precision increases from .81 to .97 compared to the GPT-4o chatbot. This increase indicates that the DCR method proposed in Study 2 is more effective in accurately extracting and coding relevant information. The total recall for DCR is slightly lower than for GPT-4o (0.98 vs. 0.99), a very small difference that suggests GPT-4o has robust performance at minimizing missed extractions.

Breaking these down further by data type, both qualitative and quantitative data extraction and coding benefited from the DCR method proposed in Study 2. As displayed in Figure 5, the precision and F1-score for qualitative data increase from .91 to .98, and from .95 to .97, respectively. Meanwhile, the most notable improvement occurs in quantitative data extraction and coding, which has been a major challenge in Study 1. The precision and F1-score of quantitative

data extraction and coding rose from .68 to .96, and from .81 to .98, respectively. These findings indicate that the DCR method in Study 2 is particularly effective in handling content-specific, less structured, and complex quantitative data with greater accuracy and completeness.

Specifically, Figure 6 presents a comparison of extraction and coding performance between the GPT-4o chatbot and the DCR method across 13 variables. Both methods exhibit consistently high performance in qualitative data extraction, with F1-scores exceeding 0.90. These variables are mostly categorical, which likely contributed to their stable extraction accuracy across models. GPT-4o chatbox achieves slightly higher F1-scores for the extraction of the type of publication and the assessment of BMI.

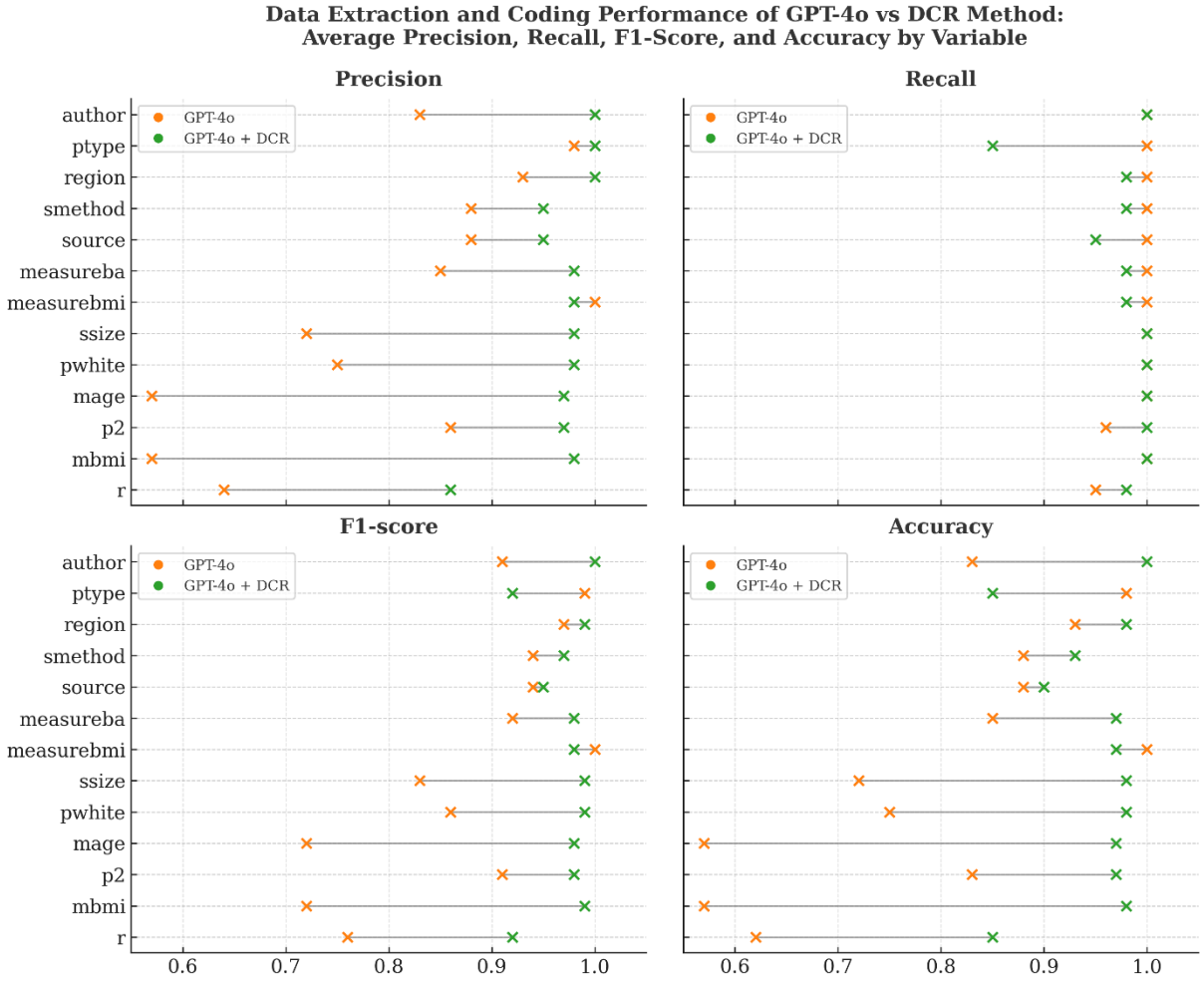

Figure 6: Data Extraction and Coding Performance of GPT-4o vs DCR Method: Precision, Recall, F1-Score, and Accuracy by Variable.

In terms of the quantitative variables, the DCR method improves the performance of extracting and coding semi-structured variables (e.g., sample size, percentage of white participants, and percentage of college students or higher), while the most substantial improvements emerge in the extraction of complex variables, particularly mean age, mean BMI, and correlation coefficients. These variables present notable challenges due to their variability in format and contextual

embedding within texts. Among them, mean age and mean BMI show the greatest performance gain, with the DCR method in Study 2 achieving approximately a 70% improvement over that of the best-performing model in Study 1. This improvement reflects the effectiveness of the DCR method in both reducing missing and inconsistent data and enhancing extraction precision.

4. Discussion

This study aims to evaluate the capabilities of LLMs in automating data extraction and coding for ES, in the context of social and behavioral science research. Building upon an existing meta-analysis, we extracted data from the papers identified by the meta-analysis study and conducted a two-step investigation: Study 1 compared the performance of different LLMs (i.e., GPT-3.5, GPT-4, GPT-4o) using manual chatbot querying, while Study 2 proposed and evaluated a novel automated method, DCR, to further optimize the extraction and validation process.

4.1. Implications of LLMs' performance

Our study provides preliminary empirical evidence on the feasibility of LLMs in automating data extraction for literature coding tasks. In our study, to answer *Research Question (1),* we compare the accuracy of three LLM-based models (i.e., GPT-3.5, GPT-4, GPT-4o) in Study 1. The results demonstrate that GPT-4o outperforms for most variables when compared with the other three models. This suggests that the extraction capabilities of LLMs are significantly associated with the advancement of newer model generations. These findings align with previous research on LLM-based data extraction in ES (Polak & Morgan, 2024; Polak et al., 2024; Wang & Luo, 2024), which suggests that LLMs are able to assist in labor-intensive data extraction tasks by reducing the reliance on manual data extraction and validation. Thus, the findings highlight LLMs' potential to significantly reduce the manual effort required for extracting data from large volumes of literature.

While LLMs show satisfactory accuracy in certain structured variables (e.g., author, assessment of BA, sample size) in Study 1, their performance (e.g., mean age, mean BMI, correlation coefficient) remains an area for further improvement for more variables. This limitation suggests that relying solely on the current LLMs for data extraction remains insufficient.

4.2. Implications of the DCR Method

Corresponding to *Research Question (2)*, the results of the DCR method proposed in Study 2 substantially improve precision and F1-score across most variables compared to that of the best-performing model in Study 1, confirming its effectiveness in data extraction and coding process of ES research with great accuracy and completeness. These improvements can be attributed to several key innovations informed by limitations identified in Study 1. First, different from previous studies that rely solely on converting into text (e.g., Khraisha et al., 2024; Wang & Luo, 2024) or focusing on converting HTML tables into customized formats (e.g., JSON; Yi et al., 2024), the DCR addresses token limitations and preserves formatting integrity by dividing entire articles into structured text sections, preserving images, and standardized Markdown format.

Second, to improve performance on complex variables such as correlation coefficients (*r*), we implemented a few-shot prompting strategy, drawing on prior approaches (e.g., Ekuma, 2024; Yi et al., 2024), but with a key difference that our prompts incorporate self-reflective feedback. By manually retrieving failed examples, providing annotated feedback, and training the model to learn from mistakes, the performance of correlation coefficient in Study 2 (precision = .86, recall = .98, and F1-score = .92) significantly improved that addresses one of the most complex variables

in the best-performing model in Study 1 (precision = .64, recall = .95, and F1-score = .76). This approach enables the model to better extract embedded statistical statements and values from diverse table formats.

Third, to address *Research Question (3),* the DCR method implements a systematic validation and rechecking process to handle missing and inconsistent data. This issue is frequently observed in Study 1 and in prior research (Garthlehner et al., 2024; Wang & Luo, 2024). Unlike MetaMate (Wang & Luo, 2024), which addresses only missing data, our method also corrects inconsistencies, making it more comprehensive. By logging validation outcomes and prompting GPT-4o to revisit discrepancies, this approach not only increased accuracy but also reduced the need for extensive human-in-the-loop (HITL) involvement.

In summary, our proposed DCR improves automation, accuracy with reliability, and traceability in LLM-assisted ES research. However, given the occasional errors and inconsistencies observed, human oversight remains crucial, particularly in tasks requiring nuanced interpretation across papers with diverse formats and varying structures. Future work should focus on model fine-tuning, further refining prompt strategies targeting specific contexts (e.g., demographic groups), and developing best practices to reduce errors and increase generalizability in large-scale ES studies.

4.3. Limitations and Future Directions

While this study demonstrates the potential of LLMs in processing automation in ES, particularly in data extraction and validation processes, several limitations still need to be noted. First, the extraction performance was not consistently stable across all variables or models. For example, performance on the type of publication and the assessment of BMI slightly declined in Study 2. The potential reason might be the information lost in the file processing process (e.g., the conversion from PDF articles into Markdown omitted the headnote and footnote in this study), or the inherent variability of LLM output. Thus, a better file processing programme with less information lost and further refined prompts with specific domain requirements are needed. A similar discussion about prompt engineering was claimed by Khraisha (2023) that a reliable prompt would raise GPT-4's screening performance to be almost perfect.

Second, the black-box issue in LLM research, where the underlying mechanisms driving outputs are not fully transparent (Komera & Manche, 2023), makes it difficult to further diagnose failure or improve extraction accuracy. Thus, the development of interpretability tools is needed to provide the model's inner decision-making processes (Caruana et al., 2015; Komera & Manche, 2023).

Third, although numerous other LLMs, such as Claude, Gemini, and DeepSeek, are available, we selected models only embedded in ChatGPT for this study. This choice was made to reduce variation across systems: models from different vendors might use different architectures, training data, and inference settings, which can confound comparisons. Notably, however, non-ChatGPT models may perform as well as or better than the models studied here. Ntinopoulos and colleagues (2025) have demonstrated that models like Claude 3.0 Opus, Gemini Advanced, and Llama 3-70b exhibit the best performance in entity extraction and binary classification tasks. These findings suggest that expanding model selection beyond ChatGPT-based models could enhance the generalizability and robustness of automated data extraction. Future research should further explore the effectiveness and performance of these additional LLMs for automated data extraction in the ES process.

Beyond data extraction, LLMs could play a broader role in other stages of ES. For example, Susnjak (2024) demonstrated the potential of LLMs in automating knowledge synthesis, a critical component of systematic literature reviews. Within the research, it suggests that future research should explore the integration of LLMs into more advanced stages of ES, such as summarization and integration of findings, while still addressing the current limitations in data extraction.

Another critical challenge is the phenomenon of hallucination, where LLMs generate fabricated or inaccurate information (Beutel et al., 2023). Although the DCR method proposed in this study significantly improves performance, hallucination remains a concern for the reliability and trustworthiness of LLM-generated outputs. For instance, LLMs may occasionally respond with "None" or provide fabricated results, which could undermine confidence in their utility for ES (Ji et al., 2023). Additionally, due to the inherent probabilistic nature of LLMs, there remains a possibility that the model might not generate consistent results across iterations of the same process (Polak and Morgan, 2024), further complicating the reproducibility and reliability. At present, maintaining the accuracy and reliability of data extraction still depends on rigorous validation processes. Besides the automated validation process, the HITL system remains essential for ensuring accountability and verifying outputs (Korema & Manche, 2023). Therefore, continued development of interpretability tools to ensure transparency and improvement of the automated validation process in the DCR method will be vital to reduce reliance on human oversight in the implication of LLMs in ES in the long run.

5. Conclusions

This study explores the use of large language models (LLMs) to automate data extraction and coding for ES research in the context of social and behavioral research. In Study 1, we compared multiple LLMs for their performance and found that GPT-4o models achieved the highest accuracy in retrieving the targeted information. In Study 2, we proposed an automated data extraction method, CDR, which further improved the efficiency and accuracy of data extraction and coding. Our findings showed that LLMs prove promising in assisting data extraction and classification and reducing manual effort and human error. However, it is still essential to remain human-in-the-loop to ensure accuracy and reliability in practice. Future research should focus on refining method performance and establishing clear guidelines for their ethical and responsible use of LLMs for ES.

Appendix A: Prompts for Study 1

"Show me this paper's author (s) and year, Publication type (Journal or Dissertation), Country, Sample (participants numbers), Survey method (interview, paper-and-pencil, online, or mixed), Percent of white in participants, Mean age of participants, Percent of college or higher of participants, Mean BMI of participants, Source of sample (community, college, primary/secondary/high school, or mixed), Assessment of BA, Assessment of BMI (self-reported or measured), effect size, correlation between body appreciation and BMI."

Appendix B: Prompts for Study 2

Prompts for *contextual information*:

"Please extract the following values from the provided text {IMAGE_ATTACHED}. Each value has a definition to guide the extraction process. If you are unable to determine the value from the text, respond with "None."

"""Author(s), Year, Publication type (Journal or Dissertation), Country, Male Sample Size, Female Sample Size, Survey method (interview, paper-and-pencil, online, or mixed), Percent of white male participants, Percent of white female participants, Mean age of male participants, Mean age of female participants, Percent of college or higher male participants, Percent of college or higher female participants, Mean BMI of male participants, Mean BMI of female participants, Source of sample (community, college, primary/secondary/ high school, or mixed), Assessment of BA, Assessment of BMI (self-reported or measured), Male effect size, Female effect size, Correlation between male body appreciation and BMI, Correlation between female body appreciation and BMI, Confidence Level (low, moderate, high), Confidence Percentage (scale of 0% to 100%)."""

Provided text:

"""{TEXT}"""

Prompts for *specific user questions*:

"The following values should be examined and reported with gender differences if available; if not, report the data without separation by gender: sample size, Percent of white in participants, mean age of participants, Percent of college or higher of participants, Mean BMI of participants, correlation between body appreciation and BMI.

Definition for the values that we will extract today:

Gender: whether part or all of the data is reported by gender (e.g., female(s) or male(s); women or men; feminine or masculine; or other descriptions that indicate female and male). If reported by gender, specify the genders; otherwise, respond with "T".

Author: The name(s) and the author(s) of the study or paper. It should only appear after the main title of the paper or study, not in the abstract or introduction or other sections.

Year: The year of the study or paper that was published.

Publication type: The type of publication means (i.e., journal articles or dissertation). It must be one of the following two options: journal articles or dissertation.

Country: The country where the study was conducted or where the participants were recruited (i.e., North America, Europe, South America, Asia).

Sample: Total numbers of participants of the study.

Survey method: The method used to collect data (i.e., paper-and-pencil or online). It must be one of the following two options: paper-and-pencil or online.

Percent of white in participants: The percentage (%) of participants identifying as white. The text might not straifghtforwardly mention "white" but you can infer it from the context (e.g., "Caucasian", "European", etc.).

Mean age of participants: The average age of the participants in the study.

Percentage of college or higher of participants: The percentage (%) of participants with college or higher education level than college. The text{IMAGE_ATTACHED} might not straifghtforwardly mention "college" or other education level higher than "college" but you can infer it from the context (e.g., the Source of Sample might indicate "all participants were college/university students" for 100%, "all participants were primary/secondary/high school students" for 0%, etc.).

Mean BMI of participants: The average body mass index (BMI) of the participants.

Source of sample: how or where the participants were recruited (i.e., community, college, or primary/secondary/ high school). It must be one of the following three options: community, college, or primary/secondary/ high school.

Assessment of BA: The method used to assess body appreciation (BA; i.e., BAS or BAS-2; BAS stands for body appreciation scale). It must be one of the following two options: BAS or BAS-2.

Assessment of BMI: The method used to assess BMI (i.e., self-reported or measured). It must be one of the following two options: self-reported or measured.

Correlation (r) between body appreciation and BMI: The reported correlation coefficient (e.g., Pearson's r) between Body Appreciation and BMI.

Confidence Level (low, moderate, high): Please also tell me, how would you rate your confidence (low, moderate, high) in your each of the answer you provided.

Confidence Percentage (scale of 0% to 100%): on a scale of 0% to 100%, how confident are you in the accuracy of each of the answer that you provided.

Examples of prompts with few-shot learning:

"Here is an example of where we should extract the correlation coefficient. The answer is that the correlation between BAS-2 and BMI for men is –0.24, the correlation between BAS-2 and BMI for women is 0.39. Please indicate and learn why this is the case and please reflect in the prompt conditions on what you should do for troubleshooting."

After manually adjusted for the self-reflection prompt suggestions:

"Correlation between body appreciation and BMI: The reported correlation coefficient (e.g., Pearson's r) between Body Appreciation and BMI, adhering strictly to the following instructions:

---1. General Extraction: Identify the reported correlation coefficient between Body Appreciation and BMI. It must be a numerical value (e.g., 0.5, -0.3, etc.). Pay attention to the table titles or notes if the information is presented in tables or matrices.

---2. Case Handling:

---Case 1 (Multiple Subscales): If multiple BAS or BAS-2 subscales correlations with BMI are reported, compute and report the average correlation across these subscales. If only one overall correlation (no subscales) is reported, directly report this single correlation.

---Case 2 (Gender-Specific Data): If correlations are separately reported by gender: For each gender, compute and report the average correlation across BAS or BAS-2 subscales if multiple subscales are present. For example, if the BAS subscales "body valorisation" and "body care" or other subscales are reported separately for males and females, or BAS scales or subscales are reported separately for heterosexual and non-heterosexual females or males, compute the average correlation for each gender. If only a single correlation per gender (no subscales) is reported, directly report this correlation for each gender.

---Case 3 (Correlations Presented in Tables or Matrices): If correlations appear in tables or matrices: Carefully examine table title and notes or matrix entries to determine whether correlations are presented by gender (e.g., separate correlations for males and females indicated by upper/lower triangles). If gender-specific correlations are indicated, separately extract and report correlations per gender following instructions in Case 2. If no gender distinction is present, directly extract and report the correlation(s) according to instructions in Case 1. There may be multiple tables or images to present the correlation, so you need to check all of them and calculate the average correlation across all the tables or images."

Prompts for *structured response format*:

***IMPORTANT:** Replace {RESULT} with the value you found or "None" if you cannot find anything*

***IMPORTANT:** If information is not explicitly provided in the text, answer with "None"*

***IMPORTANT:** Provide only the values requested, formatted exactly as specified. Any additional text will be considered an error.*

Author(s): {RESULT}

Year: {RESULT}

Publication type (Journal or Dissertation): {RESULT}

Country: {RESULT}

Male Sample Size: {RESULT}

Female Sample Size: {RESULT}

Survey method (interview, paper-and-pencil, online, or mixed): {RESULT}

Percent of white male participants: {RESULT}

Percent of white female participants: {RESULT}

Mean age of male participants: {RESULT}

Mean age of female participants: {RESULT}

Percent of college or higher male participants: {RESULT}

Percent of college or higher female participants: {RESULT}

Mean BMI of male participants: {RESULT}

Mean BMI of female participants: {RESULT}

Source of sample (community, college, primary/secondary/ high school, or mixed): {RESULT}

Assessment of BA: {RESULT}

Assessment of BMI (self-reported or measured): {RESULT}

Male effect size: {RESULT}

Female effect size: {RESULT}

Correlation between male body appreciation and BMI: {RESULT}

Correlation between female body appreciation and BMI: {RESULT}

Confidence Level (low, moderate, high): {RESULT}

Confidence Percentage (scale of 0% to 100%): {RESULT}

Appendix C:

General Prompt Template for Structured Data Extraction in Social and Behavioral Science

***Prompts for contextual information*:**

"Please extract the following values from the provided text and images ***"""{IMAGE_ATTACHED_1}"""*****.** *Each value has a definition to guide the extraction process. If you are unable to determine the value from the text, respond with "None."*

Provided text:

"""{TEXT_1}"""

All Variables for Extraction: ***"""{PASTE ALL YOUR INTERESTED VARIABLES AND THEIR CODING SCHEMES HERE}"""***

***Prompts for specific user questions*:**

"The following values should be examined and reported with ***"""{INSERT THE GROUP DIFFERENCE YOU INTERESTED, e.g., sex}"""*** *if available; if not, report the data without separation by* ***"""{GROUP DIFFERENCE}"""*****.**

Here are variables with ***"""{GROUP DIFFERENCE}: {INSERT VARIABLES WITH GROUP DIFFERENCE} """***

Definition for the values that we will extract today:

"""{VARIABLE X (Categorical data): Definition and Description;

VARIABLE Y (Continuous data): Definition and Description}"""

Prompts with few-shot learning:

Here is an example text of where we should extract ***"""{VARIABLE Y}"""****.*

Example text or images:

"""{TEXT_2}""" or
"""{IMAGE_ATTACHED_2}"""

The answer is that ***"""{INSERT THE GROUND TRUTH OF VARIABLE Y}"""****. Please indicate and learn why this is the case and please reflect in the prompt conditions on what you should do for troubleshooting."*

After manually adjusting for the self-reflection prompt suggestions:

"""{INSERT THE SELF-REFLECTION PROMPTS HERE }"""

Prompts for *structured response format*:

***IMPORTANT:** Replace {RESULT} with the value you found or "None" if you cannot find anything*

***IMPORTANT:** If information is not explicitly provided in the text, answer with "None"*

***IMPORTANT:** Provide only the values requested, formatted exactly as specified. Any additional text will be considered an error.*

"""{VARIABLE X: {RESULT}

VARIABLE Y: {RESULT}}"""

Reference

Ades, A. E., & Sutton, A. J. (2006). Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *Journal of the Royal Statistical Society Series A: Statistics in Society, 169*(1), 5-35. https://doi.org/10.1111/j.1467-985X.2005.00377.x

Beutel, G., Geerits, E., and Kielstein, J. T. (2023). Artificial hallucination: Gpt on lsd? *Critical Care, 27*(1):148.

Blaizot, A., Veettil, S. K., Saidoung, P., Moreno-Garcia, C. F., Wiratunga, N., Aceves-Martins, M., ... & Chaiyakunapruk, N. (2022). Using artificial intelligence methods for systematic review in health sciences: A systematic review. *Research Synthesis Methods*, *13*(3), 353-362. https://doi.org/10.1002/jrsm.1553

Briner, R. B., Denyer, D., & Rousseau, D. M. (2012). Systematic review and evidence synthesis as a practice and scholarship tool. In *Oxford Handbook of Evidence-Based Management* (pp. 1–2). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199763986.013.0007

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv*:2303.12712. https://doi.org/10.48550/arXiv.2303.12712

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *In Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1721-1730).

Celikten, T., & Onan, A. (2025). Benchmarking Large Language Models for Biomedical Literature Summarization: Abstractive vs. Extractive Paradigms. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3604351

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., ... & Jain, A. (2024). Structured information extraction from scientific text with large language models. *Nature communications, 15*(1), 1418. https://doi.org/10.1038/s41467-024-45563-x

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G. P., Ferragina, P., Tozzi, A. E., & Rizzo, C. (2023). ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers in Public Health, 11,* 1166120. https://doi.org/10.3389/fpubh.2023.1166120

Edwards, K. M., Song, B., Porciello, J., Engelbert, M., Huang, C., & Ahmed, F. (2023). ADVISE: AI-accelerated design of evidence synthesis for global development. *arXiv preprint arXiv*:2305.01145. https://doi.org/10.48550/arXiv.2305.01145

Ekuma, C. (2024). Dynamic in-context learning with conversational models for data extraction and materials property prediction. *arXiv preprint arXiv*:2405.10448. https://doi.org/10.48550/arXiv.2405.10448

Gartlehner, G., Kahwati, L., Hilscher, R., Thomas, I., Kugley, S., Crotty, K., Viswanathan, M., Nussbaumer-Streit, B., Booth, G., Erskine, N., Konet, A., & Chew, R. (2024). Data

extraction for evidence synthesis using a large language model: A proof-of-concept study. *Research Synthesis Methods*, *15*(4), 576–589. https://doi.org/10.1002/jrsm.1710

He, J., Sun, S., Lin, Z., & Fan, X. (2020). The association between body appreciation and body mass index among males and females: A meta-analysis. *Body Image, 34*, 10-26. https://doi.org/10.1016/j.bodyim.2020.03.006

Hill, J. E., Harris, C., & Clegg, A. (2024). Methods for using Bing's AI-powered search engine for data extraction for a systematic review. *Research Synthesis Methods*, *15*(2), 347-353. https://doi.org/10.1002/jrsm.1689

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), 1–38. https://doi.org/10.1145/3571730

Khan, K. S., Kunz, R., Kleijnen, J., & Antes, G. (2003). Systematic reviews to support evidence-based medicine: How to review and apply findings of healthcare research. *Royal Society of Medicine.*

Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., & Hadfield, K. (2024). Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods, 15*(4), 616–626. https://doi.org/10.1002/jrsm.1715

Kim, H., Zhang, X., Han, Y., He, J., & Ji, F., (2024). Assessing ChatGPT as a power analysis tool: An empirical investigation. PsyArXiv Preprint. https://doi.org/10.31234/osf.io/32mkv.

Komera, O., & Manche, R. (2023). Black-Box Behavior in Large Language Models: Challenges and Implications.

Littell, J. H., Corcoran J., & Pillai, V. (2008). Systematic reviews and meta-analysis. *Oxford University Press.*

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics, 12,* 157-173. https://doi.org/10.1162/tacl_a_00638

Mammides, C., & Papadopoulos, H. (2024). The role of large language models in interdisciplinary research: Opportunities, challenges and ways forward. *Methods in Ecology and Evolution. https://doi.org/10.1111/2041-210x.14398*

Mathes, T., Klaßen, P., & Pieper, D. (2017). Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. *BMC Medical Research Methodology, 17,* 1-8. https://doi.org/10.1186/s12874-017-0431-4

Ntinopoulos, V., Biefer, H. R. C., Tudorache, I., Papadopoulos, N., Odavic, D., Risteski, P., ... & Dzemali, O. (2025). Large language models for data extraction from unstructured and semi-structured electronic health records: a multiple model performance evaluation. *BMJ Health & Care Informatics, 32*(1), e101139.

Ohlsson, A. (1994). Systematic reviews—theory and practice. *Scandinavian Journal of Clinical and Laboratory Investigation, 54,* 25–32. https://doi.org/10.3109/00365519409088573

OpenAI(2023). GPT-4 technical report. *ArXiv preprint arXiv:2303.08774*. https://doi.org/10.48550/arXiv.2303.08774

Paruchuri, A., Garrison, J., Liao, S., Hernandez, J., Sunshine, J., Althoff, T., ... & McDuff, D. (2024). What are the odds? Language models are capable of probabilistic reasoning. *arXiv preprint arXiv:*2406.12830. https://doi.org/10.48550/arXiv.2406.12830

Polak, M. P., Modi, S., Latosinska, A., Zhang, J., Wang, C. W., Wang, S., ... & Morgan, D. (2024). Flexible, model-agnostic method for materials data extraction from text using general purpose language models. *Digital Discovery, 3*(6), 1221-1235. https://doi.org/10.1039/D4DD00016A

Polak, M. P., & Morgan, D. (2024). Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications, 15*(1), 1569. https://doi.org/10.1038/s41467-024-45914-8

Sanbonmatsu, D. M., Cooley, E. H., & Butner, J. E. (2021). The Impact of Complexity on Methods and Findings in Psychological Science. *Frontiers in Psychology*, *11*, 580111. https://doi.org/10.3389/fpsyg.2020.580111

Su, J., Jiang, C., Jin, X., Qiao, Y., Xiao, T., , H., Wei, R., Jing, Z., Xu, J., & Lin, J. (2024). Large Language models for forecasting and anomaly detection: A systematic literature review. *ArXiv,* abs/2402.10350. https://doi.org/10.48550/arXiv.2402.10350

Sun, Z., Zhang, R., Doi, S. A., Furuya-Kanamori, L., Yu, T., Lin, L., & Xu, C. (2024). How good are large language models for automated data extraction from randomized trials?. *medRxiv*, 2024-02. https://doi.org/10.1101/2024.02.20.24303083

Susnjak, T. (2024). Applying bert and chatgpt for sentiment analysis of lyme disease in scientific literature. In *Borrelia burgdorferi: Methods and Protocols* (pp. 173-183). New York, NY: *Springer US*. https://doi.org/10.1007/978-1-0716-3561-2_14

Vidal Perez, C. (2024). Prompt engineering for filling in evidence tables (Master's thesis).

Wang, X., & Luo, G. (2024). Metamate: Large language model to the rescue of automated data extraction for educational systematic reviews and meta-analyses. *Society for Research on Educational Effectiveness.* https://doi.org/10.35542/osf.io/wn3cd

Wang, S., Scells, H., Zhuang, S., Potthast, M., Koopman, B., & Zuccon, G. (2024). Zero-shot generative large language models for systematic review screening automation. In *European Conference on Information Retrieval* (pp. 403-420). *Springer, Cham*. https://doi.org/10.1007/978-3-031-56027-9_25

Yi, G. H., Choi, J., Song, H., Miano, O., Choi, J., Bang, K., ... & Kim, D. (2024). MaTableGPT: GPT-based table data extractor from materials science literature. *arXiv preprint arXiv*:2406.05431. https://doi.org/10.48550/arXiv.2406.05431

Zhang, X., Chowdhury, R., Gupta, R., & Shang, J. (2024). Large language models for time series: A survey. *ArXiv,* abs/2402.01801. https://doi.org/10.48550/arXiv.2402.01801