Journal of Academic Writing
Vol. 15 No. S2 2025, pages 1-20
https://doi.org/10.18552/joaw.v15iS2.1129

Reliability of Large Language Models for Identifying and Classifying Content in Research Articles

Kristin I. Terrill
Iowa State University, United States

Elena Cotos
Iowa State University, United States

Abstract

GenAI has demonstrated functionality that seems, uncannily, to parallel reading and writing by identifying/reformulating information from source texts and generating novel content and argumentation. These skills are essential yet challenging for many students tasked with producing literature reviews. This study takes the first steps to investigating the feasibility of a GenAI-facilitated literature review. This investigation starts from the 'human-in-the-loop' position that complex processes can be deconstructed and compartmentalized, and that component functions needed for these processes can be delegated to machines while humans contribute to, or control, the overall process. We explore the hypothesis that certain functions of the literature review process, such as information extraction and content classification, might be able to be automated. Prompts modeled on recommended practices for research synthesis were designed to identify and classify particular types of content in research articles. Outputs produced by two GenAI models, GPT-3.5 and GPT-4o, were assessed for reliability with a human coder. Overall, the results posit concerns about the models' performance on this task, cautioning against direct uses of GenAI output as learning scaffolding for students developing literature review skills.

Introduction

The generative artificial intelligence (GenAI) and large language model (LLM)[1] innovations of the past half decade, and the past two years in particular, have evoked a whirlwind of interest among researchers and teachers of written communication. In academia and education, these technologies are being tested as personal tutors and assistants for learners (Imran & Almusharraf, 2023; Bedington et al., 2024), and there is a growing interest in the affordances of LLMs as tools for expediting scholarly processes including reviewing literature. A critical position has emerged toward such uses, centering on the technical limitations of LLMs and encompassing surrounding ethical questions. Meanwhile, numerous products are emerging with claims that they facilitate literature review by integrating LLMs and scholarly literature datasets into bespoke computer applications (e.g., Scite, SciSpace, Elicit, Consensus). This

[1] Throughout this paper, GenAI is used as a general term to mean artificial intelligence (AI) systems that can output content, including text, graphics, audio, and other formats, in response to inputted prompts. LLM is used to mean a class of AI models designed to perform language-related tasks, capable of creating human-like text based on provided input. AI is used to refer generically to algorithms that utilize large datasets and statistical analysis for a range of purposes, including LLMs, GenAI, natural language processing and machine learning. For a concise discussion of these terms in everyday usage, see Toner (2023).

study engages with these issues and considers the prospect of a partially LLM-assisted literature review process. As a starting point, we focus on human–computer reliability, a fundamental prerequisite for using LLM outputs as a scaffold for developing novice scholars' literature review skills.

Arguably, reviewing literature is an essential aspect of scientific research by which scholars gain an understanding of the state of knowledge about a topic and situate their contributions meaningfully for their peers and posterity. However, it is also an intensively laborious process, as it requires time and considerable human effort to find, select, read, and synthesize appropriate and credible sources to then develop coherent arguments and generate new knowledge. This inherent constraint may obstruct the progress of scientific discovery (Hope et al., 2023). Therefore, among the proposed applications of LLMs, the idea of using them to expedite the literature review process has gained momentum among scholars in diverse disciplines (Khalifa & Albadawy, 2024; Wagner et al., 2022), paving the way for automating component functions of literature review (Alshami et al., 2023; Álvarez-Martínez et al., 2023; Ngwenyama & Rowe, 2024; Khraisha et al., 2024; Susnjak et al., 2024). Early forays with LLMs have revealed that, while the overall time spent on search, literature selection, information extraction, and knowledge synthesis can be reduced by employing AI (Wagner et al., 2022), LLMs' inaccuracies necessitate human checking (Alshami et al., 2023; Kacena et al., 2023).

That said, the current need for human oversight does not necessarily negate the potential utility of LLMs. Although fully autonomous algorithms are sometimes idealized (Susnjak et al., 2024), other theorists frame human–computer interaction as a means of extending human intellectual abilities beyond the natural limit (Tang, 2020). Theorists envision AI-integrated workflows that draw on humans' creative and reasoning abilities while exceeding human information-processing capacity (Khalifa & Albadawy, 2024; Knowles, 2024; Wagner et al., 2022). This idea echoes the concept of *scaffolding* in educational psychology, whereby learners unable to independently complete a complex task develop component skills when a tutor helps them with aspects beyond their abilities (Wood et al., 1976). Literature review poses distinct challenges to novice scholars (Carver et al., 2013; Chen et al., 2016), beyond being laborious. Therefore, the present study is motivated by the assumption that LLM affordances could be leveraged to support them in this process.

Research on implementing LLMs as literature review aids is generally concerned with the methodology for systematic reviews and meta-analyses, without addressing developmental issues that pertain to novice scholarly writers. Meanwhile, writing scholarship related to LLMs abounds, exemplified in special issues of *Computers and Composition* (vol. 71/March 2024), *Double Helix* (vol. 11), and this issue of the *Journal of Academic Writing*. This burgeoning domain of inquiry speaks to the urgency of establishing a knowledge base in writing-focused pedagogical affordances and applications of LLMs. However, at present, the scholarship has been largely theoretical (e.g., Knowles, 2024) or concentrated on learners' explorations and experiences (e.g., Bedington et al., 2024) without testing LLMs' suitability for specific teaching and learning tasks.

Literature review draws on literacy skills (Kim, 2020). Boell and Cecez-Kecmanovic (2014) put forth a literature review framework, where reading skills undergird numerous component functions. In their framework, reading is a crucial step in the process that acts as a swivel between hermeneutic circles and cycles. The first circle encompasses functions such as searching, sorting, and selecting articles; the second circle necessitates a higher degree of human reasoning and creativity for mapping and classifying content and critically assessing the literature in light of available knowledge. On the other hand, undeveloped reading skills have been theorized as a root of challenges that novices may encounter when conducting a literature review (Chen et al., 2016). Thus, scaffolding reading for novice literature reviewers could be a beneficial leveraging of LLM affordances. Research investigating LLMs' ability to simulate human reading skills suggests that they are a major step forward from alternative natural language processing approaches (Chen et al., 2023; Hoffman et al., 2021; Miao et al., 2019; OpenAI, 2023a). Yet, Miao et al. (2019) reported inconsistencies in LLMs' performance depending on how information is presented, with better reliability at identifying explicitly presented facts but less impressive identification of implicit information, intent, and sentiment.

LLMs are novel and their reliability is unproven, and they are already being promoted as tools for scientific research. Expediting and enhancing human researchers' abilities to interpret scholarly literature forms the basis of marketing for LLM-integrating software products that promise to "analyze research at superhuman speed" (Elicit, 2024) and "give researchers unmatched insight into any topic" (Scite, 2024). These tools may layer multiple types of algorithms, such as search, machine learning, and non-LLM natural language processing. Documentation on how LLMs fit into the overall configuration of these applications can be sparse or hard to find. In this way, the quality or faultiness of these products' outputs is left to the users to trace and attribute. It is worthwhile to establish clarity about the abilities and limitations of LLMs, themselves, to enable evaluation of complex and opaque systems that integrate them.

With this in mind, the present study investigates LLMs' reliability at identifying and classifying content in research articles, which is directly related to a key component function requiring reading comprehension in the literature review process (Bernhardt, 2023; Boell & Cecez-Kecmanovic, 2014). Among the many processes that comprise a literature review, identifying and classifying content were our focus since, first, they relate to information processing, and second, they constitute the functionality of computer applications currently being marketed as literature review assistants. Other necessary processes involved in conducting a literature review, such as synthesizing information and developing an argument, rely more on human creativity and reasoning, and fall outside the scope of the current study. Our study design involves human and LLM analysis of peer-reviewed articles. A human reader coded text excerpts into 11 content categories, which were drawn from schemas recommended in previous research and targeted by LLM-integrating tools (e.g., SciSpace, Elicit, Consensus), also accounting for the manner in which information was presented. Human coding was used as a standard for assessing the reliability of two LLMs: GPT-3.5 Turbo (OpenAI, 2023b) and GPT-4o (OpenAI, 2024). To test whether the functions of *content identification* and *classification* in the literature review process can be reliably delegated to LLM, GPT-3.5 and GPT-4o were prompted to classify the information of interest into the same content categories. The LLM outputs were compared to the coder's classifications, and human–computer reliability was calculated. The results revealed that the reliability was not consistent for either LLM and that both LLMs were less reliable when it came to implicitly presented information that requires reader inferencing.

Methods

Research article compilation

To begin with, we compiled research articles in PDF format. We used 'ARTIFICIAL INTELLIGENCE' and 'ACADEMIC WRITING' as search terms in Scopus. The 71 search results were reviewed, and we used three criteria—(a) topic relevance, operationalized as investigating LLMs and academic writing; (b) research study with empirical results; and (c) English as the language of publication—to select five representative articles (see Table 1), as this number allowed for the generation of 50 comparisons within each content category (see 'Content categories' under Methods).[2]

Human coding

Human coding of the sample articles entailed, first, coding texts based on content categories. The manner by which this information was presented in each article was subsequently coded as well, following the three-level schema provided in Miao et al. (2019). Coding was done by an early career, applied linguistics researcher with experience in text analysis and literature review.

[2] This exceeds the 30 comparisons recommended for estimating agreement reliability (McHugh, 2012).

Table 1. Sample Articles Manually Classified and Integrated into LLM Prompts

Article	Citation
Article 1	Chauke, T.A., Mkhize, T.R.; Methi, L., & Dlamini, N. (2024). Postgraduate students' perceptions on the benefits associated with artificial intelligence tools for academic success: The use of the ChatGPT AI tool. *Journal of Curriculum Studies Research 6(*1). https://doi.org/10.46303/jcsr.2024.4
Article 2	Danler, M., Hackl, W.O., Neururer, S.B., & Pfeifer, B. (2024). Quality and effectiveness of AI tools for students and researchers for scientific literature review and analysis. *Studies in Health Technology and Informatics 313*. https://doi.org/10.3233/SHTI240038
Article 3	Duah, J. E., & McGivern, P. (2024). How generative artificial intelligence has blurred notions of authorial identity and academic norms in higher education, necessitating clear university usage policies. *International Journal of Information and Learning Technology 41*(2). https://doi.org/10.1108/IJILT-11-2023-0213
Article 4	Johnston, H., Wells, R.F., Shanks, E.M., Boey, T., & Parsons, B.N. (2024). Student perspectives on the use of generative artificial intelligence technologies in higher education. *International Journal for Educational Integrity 20*(1). https://doi.org/10.1007/s40979-024-00149-4
Article 5	Mahapatra, S. (2024). Impact of ChatGPT on ESL students' academic writing skills: A mixed methods intervention study. *Smart Learning Environments 11*(1). https://doi.org/10.1186/s40561-024-00295-9

Content categories

Table 2 presents 11 content categories used for coding. Most of the categories were derived from Bernhardt's (2023) recommendations for literature synthesis. Bernhardt proposed nine content categories, of which six incorporate Melnyk et al.'s (2016) Level of Evidence and PICOT[3] conceptual frameworks. Two content categories—*Paradigm* and *Research Design*—were added since they are deemed necessary to prime an analyst for critical assessment (Boell & Cecez-Kecmanovic, 2014). To enable comparative analysis, the coder read and coded all five articles. This task heavily relied on reading comprehension, as the coder conveyed their understanding of the content by writing a concise description of information pertaining to a content category in their own words (see examples in Table 4).

The choice of these content categories was reinforced by their salience in commercial LLM-incorporating applications used for information extraction. For example, Consensus (2024) includes the 'Study Snapshots' feature, which extracts population, sample size, methods, and outcomes measured. For its 'Extract data' feature, Elicit (2024) recommends adding multiple categories, including main findings, intervention, and outcome measured. Elicit does not disclose the model used to generate its outputs, but Consensus states that their technology integrates GPT-4 (Consensus, 2024).

Manner of information presentation categories

Drawing on the conclusions by Miao (2019) on the reading comprehension abilities of transformers and considering that content information in articles is presented in various ways, the coder accounted for the manner of presentation; that is, whether content information was conveyed as explicitly presented information, inferenced through equivalent terms in context, or inferenced through related common knowledge. Descriptions of each manner of information presentation are provided in Table 3, and examples can be found in Results and Discussion.

LLM information extraction and content classification

To parallel the human content coding task, GenAI was used to perform an information extraction and classification task. Two LLMs were tested: GPT-3.5 and GPT-4o. GPT-3.5 is the well-known LLM that underlay the web application ChatGPT when it was released in 2022. GPT-4o is a new generation of the GPT family. It is important to assess both models' reliability as literature review assistive tools since either model may be integrated into applications students might use, such as Consensus.

[3] PICOT framework: **P**opulation of **I**nterest, **I**ntervention or **I**ssue of Interest, **C**omparison Intervention or Group, **O**utcome, and **T**imeframe (Melnyk et al., 2016).

Table 2. Content Categories

Content Category	Description
Article Type	Type of inquiry (e.g., research, review)
Purpose	Overarching idea that anchors the article
Paradigm	Epistemic position rationalizing knowledge claims
Research Design	Collection, measurement, and analysis of data
Level of Evidence	• Systematic reviews of randomized control trials • Randomized control trials • Controlled trials without randomization • Case-control and cohort studies • Systematic reviews of descriptive and qualitative studies • Single descriptive or qualitative studies • Expert opinions
Population	Features of study specimens or participants
Intervention/Issue of Interest	Experimental conditions, independent variables, or treatments
Comparison Intervention/Group	Explanation of conditions applied to the untreated control group
Timeframe	Amount of time observed
Outcome	Result of experiment
Major Findings	Inferences or deductions from data analysis

Table 3. Manner of information presentation

Information Presentation Type	Description
Explicitly Presented Information	Information was presented using the exact terminology used to label the content category
Inference through Equivalent Terms in Context	Information was presented using an equivalent term or phrase as that used to label the content category
Inference through Related Common Knowledge	Information was presented without any labeling, requiring the reader to draw on their discourse or content knowledge

Text extracted from the PDF files was integrated into a prompt (see Appendix A) to obtain LLM classification that would be comparable to the above-described human coding. Defining each content category, the prompt was developed following recommendations in Mollick (2023), such as providing context and constraints and giving explicit step-by-step instructions. It was then modified five times to integrate each of the five input texts, so that the LLMs could produce respective outputs. Due to OpenAI's input size restrictions, only the full text of the first page of each article was integrated into each prompt.[4] Using the OpenAI application programming interface (API), the prompts were iterated 10 times for each article, enabling analysis of individual article effect on the LLMs' accuracy. Since there were five prompts integrating text from five different articles, and since each prompt was given to the LLMs 10 times, 50 tests were run for GPT-3.5 and 50 tests for GPT-4o. Therefore, each LLM generated 50 outputs; an output contained information pertaining to 11 content categories (Table 1). Examples of LLM outputs are provided in Appendix B. For reliability analysis, the outputs were analyzed in terms of content category and manner of information presentation; see Figures 1 and 2, and Tables 9–11.

[4] Nearly all the recommended article synthesis information was provided on the first page of all sampled articles. When the full article text was included in the prompt, it was found to exceed the token limit for the GPT-3.5 API. For these reasons, it was deemed practical to provide only the first page of each article.

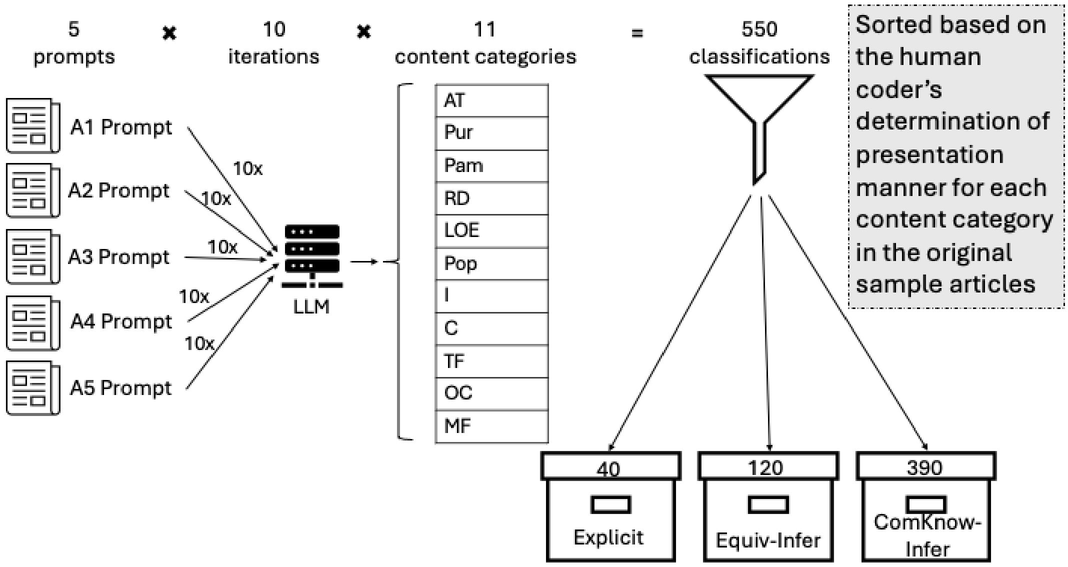
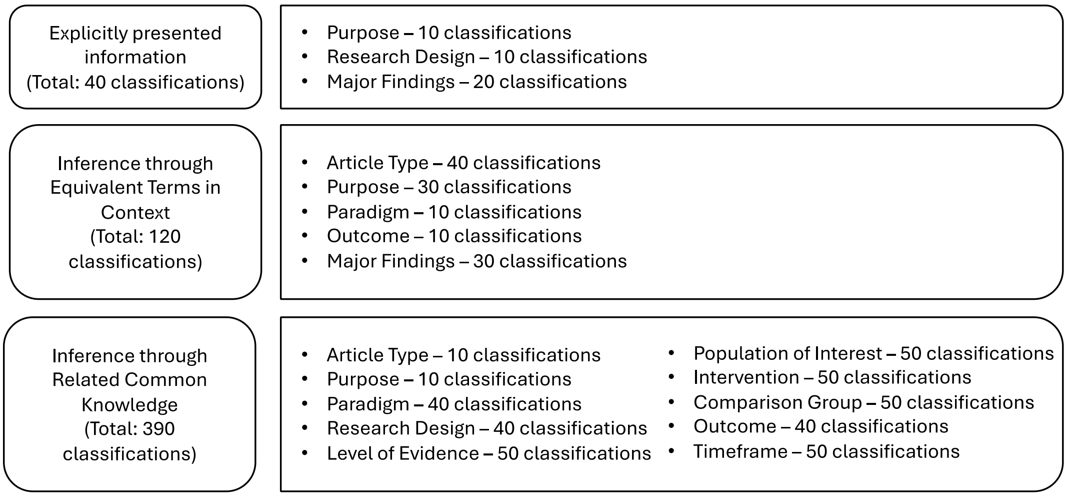


Figure 1. LLM Content Classification Workflow and Output Sorting



Explicitly presented information (Total: 40 classifications)

- Purpose – 10 classifications
- Research Design – 10 classifications
- Major Findings – 20 classifications

Inference through Equivalent Terms in Context (Total: 120 classifications)

- Article Type – 40 classifications
- Purpose – 30 classifications
- Paradigm – 10 classifications
- Outcome – 10 classifications
- Major Findings – 30 classifications

Inference through Related Common Knowledge (Total: 390 classifications)

- Article Type – 10 classifications
- Purpose – 10 classifications
- Paradigm – 40 classifications
- Research Design – 40 classifications
- Level of Evidence – 50 classifications
- Population of Interest – 50 classifications
- Intervention – 50 classifications
- Comparison Group – 50 classifications
- Outcome – 40 classifications
- Timeframe – 50 classifications

Figure 2. Classification Counts Grouped by Information Presentation Manner and Content Categories

Comparison of human coding and LLM classification

Outputs produced by both LLMs were exported to Excel spreadsheets, and the coder marked each instance of their output as either agreement or disagreement. In this process, the coder considered content as restricted or non-restricted, meaning that each content category could be determined based on the nature of the information. For example, *Paradigm* was considered restricted because the articles described either qualitative, quantitative, or mixed-methods research. Similarly, *Levels of Evidence* was restricted to seven levels, per Melnyk et al. (2016). On the other hand, content related to *Purpose* could be paraphrased by the coder and LLMs. Thus, categories with information allowing paraphrasing or summarizing were non-restrictive. Making this distinction was important when determining whether there was agreement or disagreement between human judgment based on reading comprehension and LLM classification. For restrictive content categories, only exact agreement was considered agreement. For example, if the coder assigned 'qualitative' as the *Paradigm*, an exact agreement output by the LLMs must be 'qualitative.' Non-restrictive categories required interpretation to determine whether LLM output conveyed the same meaning as the coder's text. Table 4 provides examples of agreement and disagreement in such categories. The first is an example of agreement because the core meaning is the same, even though the text written by the coder and produced by the LLM is different. In the second example, the meaning of the

LLM-generated excerpt is not the same as that conveyed by the coder; thus, this was marked as disagreement.

Table 4. Examples of Agreement and Disagreement for Non-restrictive Content Categories

Content Category	Human Coder	GPT-3.5	Agreement/ Disagreement
Purpose	Understand graduate students' perceptions of the benefits of using artificial intelligence tools, namely ChatGPT.	Explore postgraduate students' perceptions of the benefits associated with the utilisation of artificial intelligence tools, with a specific focus on ChatGPT, in their academic success at historically disadvantaged universities in South Africa.	Agree
Major Findings	Outputs of tools highly various, including non-scholarly sources. Phrasing of output varies, but content is stable within each tool. Data sourcing may be time-restricted. Selection method is not transparent.	Highlighted varying response qualities of AI tools, emphasized lack of transparency in source selection, and suggested further research on integrating diverse AI tools and assessing commercial tools.	Disagree

Reliability analysis

Human–LLM reliability was assessed by comparing the LLM outputs to human coding using Cohen’s kappa (Cohen, 1960). Kappa was calculated for each content category within each manner of information presentation. Because there were fewer than 30 classifications in some information presentation/content categories (see Figure 2), e.g., explicitly presented *Purpose*, some of these calculations included fewer than 30 comparisons. To provide additional clarity about the frequency with which the LLMs’ output corresponded with the coder’s, percent agreement was calculated as the quotient of agreement instances over total number of classifications conducted. Information presentation is one of countless complex communicative choices on the part of article authors that has been previously shown to impact the fact-identification performance of LLMs (Miao et al., 2019). At present it is not known whether other article-level dynamics, such as discipline, genre, publication language, or intended audience, could affect the reliability of LLMs. To address this, we added a Chi-square test of independence with respect to the article as an independent variable.

Results and Discussion

This section first reports the results of human coding by content classification and by manner of information presentation. Then, the results of the analysis of LLMs’ reliability are presented for each information presentation manner.

Human coding of content

The content codes assigned by the coder are shown in Table 5. In the *Purpose* and *Major Findings* categories, the coder paraphrased long phrases and clauses based on their reading comprehension of the information in articles. For these non-restrictive content categories, equivalent paraphrasing was possible, as opposed to the restrictive *Paradigm* and *Level of Evidence*. The PICOT content categories (Melnyk et al., 2016) were applicable in only one of the five sample articles.[5] Since the other articles were non-experimental, common knowledge inference was required to deduce that the PICOT categories were not applicable.

[5] We speculate that the novelty of LLMs may account for the limited amount of experimental studies that were available at the time this study was conducted.

Table 5. *Human Coding*

	Article 1	**Article 2**	**Article 3**	**Article 4**	**Article 5**
Article Type	Empirical research	Empirical research	Empirical research	Empirical research	Empirical research
Purpose	Understand graduate students' perceptions of the benefits of using artificial intelligence tools, namely ChatGPT	To evaluate AI tools' suitability for the purpose of scientific literature review	To examine students' and staff knowledge and attitudes toward AI and its use by students to aid in academic assessments	Understand students' use of, and beliefs about, ChatGPT for academic purposes	Measure impact of ChatGPT formative feedback on writing skills of undergraduate ESL students
Paradigm	Qualitative	Qualitative	Qualitative	Quantitative	Mixed methods
Research Design	Semi-structured interviews with thematic analysis	Exploratory; keyword search and output evaluation	Semi-structured interviews	Survey	Tests and focus group discussions
Level of Evidence	Single descriptive or qualitative studies	Single descriptive or qualitative studies	Single descriptive or qualitative studies	Single descriptive or qualitative studies	Not clear[5]
Population of Interest	Not applicable[6]	Not applicable	Not applicable	Not applicable	Undergraduate ESL students
Intervention/ Issue of Interest	Not applicable	Not applicable	Not applicable	Not applicable	AI-generated formative feedback

[6] As we discuss in the next section, the high frequency of 'not applicable' content contributed to high frequency of information requiring inference from common knowledge.

	Article 1	**Article 2**	**Article 3**	**Article 4**	**Article 5**
Comparison Intervention/ Group	Not applicable	Not applicable	Not applicable	Not applicable	Not clear[7]
Timeframe	Not applicable	Not applicable	Not applicable	Not applicable	Not clear[5]
Outcome	Not applicable	Not applicable	Not applicable	Not applicable	Significant positive impact
Major Findings	Graduate students believe it positively impacts their success, especially helpful with linguistic challenges. It could be conducive to graduate students' research output. They use it while conceptualizing studies, formulating research questions & problems, editing, and reviewing literature.	Outputs of tools are highly various and can be based on information from non-scholarly sources. Phrasing of output varies, but content is stable within each tool. Data sourcing may be time-restricted. Selection method is not transparent.	Three themes were identified: ambiguity in definition of academic misconduct (blurred notion of authorial integrity and authenticity); difficulties conceptualizing fair use of AI in academia; need for clear institutional policies on use in assessments.	Survey respondents had nuanced views on the use of AI for academic purposes, finding its use as a specific support or scaffold acceptable but not as a total replacement for engaging with the assignment.	Gains in writing skills in the experimental group were significantly higher than in the control group on the post-test and the delayed post-test.

[7] This information was not provided on the first page of the article, which was the only content provided to the LLM.

Human coding of manner of information presentation

It was observed during human coding that, in some cases, the exact terminology provided in Bernhardt (2023) was used to present information explicitly in the article, as illustrated for the *Research Design* category in Example 1.[8]

> Example 1. **Design/methodology/approach** – A qualitative method involving five one-to-one **semi-structured interviews** with four students and a lecturer explored the ethical and practical issues of GenAI text generation in academia. An inductive **thematic analysis** was chosen as it provided nuanced insights aligned with the study's goals. [Article 3]

In other cases, inference was necessary, as illustrated for *Purpose* in Example 2 and for *Research Design* in Example 3.

> Example 2. **The primary objective** was to categorize these tools based on their functionality and effectiveness in assisting with different research related tasks. [Article 2]

> Example 3. The sample comprised 10 postgraduate students pursuing master's degrees at two historically disadvantaged universities in South Africa, selected through purposive sampling. **Semi-structured interviews** were conducted to gather insights from the participating students. **Thematic analysis** was then employed to analyse the collected data. [Article 1]

Table 6 presents the manner of information presentation for each article and content category. Explicit stands for explicitly presented information, Equiv-Infer for inference through equivalent terms, and ComKnow-Infer for inference through related common knowledge. Most information was presented in a way that necessitated inference to meaningfully interpret it in terms of content categories. All five articles presented some of the relevant content in a way that required inference. Explicit was rare; three articles did not present any of the content categories explicitly.

Table 6. Content Categories and Information Presentation Manner

	Article 1	**Article 2**	**Article 3**	**Article 4**	**Article 5**
Article Type	Equiv-Infer	ComKnow-Infer	Equiv-Infer	Equiv-Infer	Equiv-Infer
Purpose	Equiv-Infer	Equiv-Infer	Explicit	Equiv-Infer	ComKnow-Infer
Paradigm	ComKnow-Infer	ComKnow-Infer	Equiv-Infer	ComKnow-Infer	ComKnow-Infer
Research Design	ComKnow-Infer	ComKnow-Infer	Explicit	ComKnow-Infer	ComKnow-Infer
Level of Evidence	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer
Population of Interest	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer
Intervention/Issue of interest	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer
Comparison Intervention/Group	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer
Timeframe	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer
Outcome	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer	ComKnow-Infer	Equiv-Infer
Major Findings	Equiv-Infer	Equiv-Infer	Explicit	Equiv-Infer	Explicit

[8] All examples provided are direct quotes from the sample articles that the human coder drew on to assign codes in each content category.

LLM information extraction and content classification

Both GPT-3.5 and GPT-4o were prompted to generate content classification tables for the five input texts, with 10 iterations of the prompt for each text. The outputs were designated as either agreement or disagreement with the coder. Raw agreement, disagreement, and agreement percentages for each content category are presented in Table 7. Both LLMs agreed with the coder for the majority of classifications in each content category, except for GPT-3.5 in the *Research Design* category, where agreement was below 50%.

Table 7. LLM information classification agreement by content category

Content Category	GPT-3.5 Agree	GPT-3.5 Disagree	GPT-3.5 Agree %	GPT-4o Agree	GPT-4o Disagree	GPT-4o Agree %
Article Type	50	0	80	50	0	100
Purpose	40	10	100	50	0	100
Paradigm	32	18	64	43	7	86
Research Design	23	27	46	33	17	66
Level of Evidence	38	12	76	40	10	80
Population of Interest	40	10	96	45	5	98
Intervention/Issue of Interest	41	9	84	45	5	90
Comparison Intervention/Group	50	0	80	50	0	90
Timeframe	50	0	82	50	0	90
Outcome	41	9	100	44	6	100
Major Findings	48	2	82	49	1	88

Although the content categories have semantic distinctions that a human reader is sensitive to, there is no theory that suggests these semantic differences would affect the accuracy or reliability of an LLM with information extraction and content classification. However, it has been shown that the manner of information presentation affects LLM performance (Miao et al., 2019). Our tests with GPT-3.5 and GPT-4o reinforce this finding, as shown in Table 8. The rates of agreement for both LLMs were highest for Explicit, which was the least frequently observed in the sample articles, and lowest for ComKnow-Infer, which was the most frequently observed in the sample articles. Part of the reason that inference through related common knowledge was the most frequent is that PICOT content categories were not applicable for four of the five sample articles. As we discuss in the next section, both LLMs achieved high reliability with the coder on the PICOT categories, but not in other content categories, when inference through common knowledge was required.

Table 8. LLM information classification by information presentation manner

Information Presentation Manner	GPT-3.5 Agree	GPT-3.5 Disagree	GPT-3.5 Agree %	GPT-4o Agree	GPT-4o Disagree	GPT-4o Agree %
Explicit	40	0	100	40	0	100
Equiv-Infer	110	10	92	116	4	97
ComKnow-Infer	303	87	78	343	47	88

Reliability of LLM classification

Overall, the human–LLM reliability varied from none to perfect. Kappa and percent agreement with both GPT-3.5 and GPT-4o are presented in Tables 9–11. Low kappa values were observed only for information requiring implicit inference, but inferential presentation did not always engender low reliability.

Explicitly presented information (Explicit)

When key information about the study design was explicitly stated in articles, as in Example 1 above, both LLM models achieved perfect reliability (see Table 9). For instance, in the *Research Design* category, LLMs were more reliable when the information was explicitly presented.

Table 9. Reliability for Explicit

Content Category	Number of Tests	GPT-3.5 κ	GPT-3.5 % agreement	GPT- 4o κ	GPT-4o % agreement
Purpose	10	1	100%	1	100%
Research Design	10	1	100%	1	100%
Major Findings	20	1	100%	1	100%

In other content categories, however, no such boost was apparent. For *Purpose* and *Major Findings*, the LLMs achieved almost perfect (κ=.90-.95) to perfect (κ=1) (McHugh, 2012) reliability on all 50 classifications (see Tables 9 and 10), regardless of whether the information was presented explicitly (Example 4), or using inference from equivalent terms or common knowledge (Example 5).

> Example 4. **This study examines the impact** of generative artificial intelligence (GenAI), particularly ChatGPT, on higher education (HE). [Article 3]

> Example 5. This paper presents a study **on the impact of ChatGPT as a formative feedback tool on the writing skills of undergraduate ESL students**. Since artificial intelligence-driven automated writing evaluation tools positively impact students' writing, ChatGPT, a generative artificial intelligence-propelled tool, can be expected to have a more substantial positive impact. However, very little empirical evidence regarding the impact of ChatGPT on writing is available. **The current mixed methods intervention study tried to address this gap**. [Article 5]

Inference through equivalent terms in context (Equiv-Infer)

When information was presented with equivalent terms, both LLMs exhibited almost perfect (κ=.90-.95) to perfect (κ=1) reliability with 93.3%–100% agreement (McHugh, 2012) in all but one content category, as shown in Table 10. Thirty comparisons were conducted on each LLM's classification of *Purpose* and *Major Findings*. Both LLMs achieved perfect agreement with respect to *Purpose*. The exception to the pattern is the *Outcome* category in Article 5, the only article reporting experimental design. For this category, percent agreement was middling, and kappa was nil for both LLMs. This may be because the LLMs were instructed not to identify PICOT details unless an article reported a clinical trial; thus, a common knowledge inference would seem to precede the expected inference of the equivalent term.

Table 10. Reliability for Equiv-Infer

Content category	Number of Tests	GPT-3.5 κ	GPT-3.5 % agreement	GPT-4o κ	GPT-4o % agreement
Purpose	30	1	100%	1	100%
Article Type	40	1	100%	1	100%
Paradigm	10	1	100%	1	100%
Major Findings	30	0.90	93%	0.95	97%
Outcome	10	0	20%	0	70%

The PICOT information was invariably presented inferentially. For four articles inference was necessary to conclude that the PICOT categories were not applicable (Table 4). While the LLMs frequently agreed with the coder about whether an article reported an experimental study, both LLMs incorrectly identified Article 1 as experimental and Article 5 as not, in several classifications, resulting in weak to moderate agreement within this category.

Inference through related common knowledge (ComKnow-Infer)

Agreement was weak in several other categories where the information presentation relied on inference from related common knowledge. Kappa and agreement percentage for the classifications in this group are presented in Table 11. In the 10 classifications conducted for *Purpose*, GPT-3.5 failed to identify a critical element of the study purpose, resulting in 0 agreement with the coder. In contrast, GPT-4o correctly identified the same study's purpose and achieved perfect agreement in all 10 classifications. *Paradigm*, *Research Design*, *Level of Evidence*, and *Intervention/Issue of Interest* were not well classified by either LLM, though GPT-4o's reliability surpassed GPT-3.5's. In these content categories, kappa ranged from 0–0.79, i.e., reliability was moderate at best (McHugh, 2012). The three content categories in which both LLMs exhibited perfect agreement with the coder were the PICOT categories of *Comparison Intervention/Group*, *Outcome*, and *Timeframe*. In each of these categories the correct output was ‘not applicable.’ In general, neither LLM reliably output information that required common knowledge inference, but they were capable of generating a null or ‘not applicable’ output when the information was completely absent. The ‘not applicable’ output was correct for half of all the classifications conducted in the ComKnow-Infer group. Reliability was lower when a positive response, i.e., anything other than null or ‘not applicable’, was needed.

Table 11. Reliability for ComKnow-Infer

Content category	Number of Tests	GPT-3.5 κ	GPT-3.5 % agreement	GPT-4o κ	GPT-4o % agreement
Article Type	10	1	100%	1	100%
Purpose	10	0	0%	1	100%
Paradigm	40	0.46	55%	0.79	83%
Research Design	40	0.27	33%	0.50	58%
Level of Evidence	50	0.38	76%	0	80%
Population of Interest	50	0.14	80%	0.90	90%
Intervention/Issue of Interest	50	0.53	82%	0.73	90%
Comparison Intervention/Group	50	1	100%	1	100%
Outcome	40	1	100%	1	100%
Timeframe	50	1	100%	1	100%

Length of output

A subsequent analysis was conducted to establish the grounds for a follow-up hypothesis about why the LLMs achieved higher agreement in the categories of *Purpose* and *Major Findings*. We speculated that these content categories engendered longer outputs than the others, and a word count analysis bore this out, as shown in Table 12 where bold indicates notably higher word counts. The coder and both LLMs used an average of 14–45 words to classify the articles in these two categories, compared to an average of 0–5 words in the other categories.

Table 12. Mean Word Counts

Content category	Subject Matter Expert	GPT-3.5	GPT-4o
Purpose	**14**	**18.76**	**23.76**
Article Type	2	1.7	1.14
Paradigm	1.2	1.56	1.3
Research Design	2	3.14	4.72
Level of Evidence	1	0.96	1
Major Findings	**29.6**	**30.06**	**44.68**
Population of Interest	2.2	0.56	1.98
Intervention/Issue of Interest	2.2	0.78	2.3
Comparison Intervention/Group	2.4	0	0.96
Outcome	2.2	0.3	2.34
Timeframe	2	0	0.26

Independence of text-specific factors

Finally, a Chi-square test of independence was conducted to ascertain whether text-specific factors could affect the frequency of LLM agreement with the coder. For both GPT-3.5 and GPT-4o, the Chi-square value (see Table 13) exceeded the Chi-square distribution value for α=0.05 with four degrees of freedom. Therefore, the null hypothesis that the text did not affect agreement frequency was rejected, and it was concluded that text-specific factors do weigh on the probability of LLM–human agreement, inclusive of all content categories.

Table 13. Chi-square Values for Independence of Article in Coder Agreement Frequency

LLM	χ^2
GPT-3.5	118.67
GPT-4o	54.63

Conclusions

As an exploratory step toward an LLM-scaffolded literature review for novice scholars, this study tested the hypothesis that GenAI might competently manage particular functions of the literature review process, namely identifying information and classifying content of interest in research articles. Our findings do not provide sufficient support to the feasibility of fully delegating this function to LLMs, as neither of the two tested LLMs exhibited consistently reliable agreement with the coder. Cumulatively, the results indicate that the reliability of LLM agreement with the coder ranged from none to perfect. Obtained insights reinforce the findings of Alshami et al. (2023) and Khraisha et al. (2024), who identified inaccuracies in the information extraction outputs of LLMs. Our findings suggest that information presentation, content category, and LLM-specific factors, such as their design and training method, may impact their performance in terms of reliability. It is worth noting that these factors are dynamic, distinct not only for articles reviewed but also the literature reviewer's approach.

Information presentation manner appears to be a key factor. Explicit information presentation may contribute to higher reliability. Both LLMs achieved full agreement with the coder when the information was presented explicitly, but they were less reliable when it came to inferential information, especially without the presence of equivalent terms. Reinforcing OpenAI's (2023a) claims about GPT-4's superiority to GPT-3.5, we found that the newer model achieved higher reliability than GPT-3.5. However, it still failed to achieve consistently strong levels of agreement when information was inferential. Because academic writing is highly implicit, trusting in readers' background knowledge and ability to infer implicit meanings (Biber & Gray, 2010), using LLMs to simulate reading comprehension in this register remains called into question. Future studies can investigate whether LLMs can be trained to better detect inferentially presented information using labeled examples from academic registers.

The results also reveal the role of the types of content. Both LLMs performed consistently well in identifying the article *Type*, *Purpose*, and *Major Findings*, regardless of whether that information was conveyed explicitly or inferentially. We speculate that the length of answer, which was longer for the *Purpose* and *Major Findings*, might account for the high level of agreement for these categories, but this would not explain the high agreement in the *Article Type* category. It is also worth noting that the length of the expected output and factors related to the article itself may affect the likelihood of human–LLM agreement. More research is needed to determine the detection of content in Melnyk et al.'s (2016) PICOT categories. Only one sample article used in this study was experimental, so it is warranted to further examine such texts to determine in what manner PICOT content is presented and how reliable LLMs are at identifying it.

An important limitation to note with this study is that the prompts used to obtain LLM outputs were not systematically optimized. Systematic revision of the prompt, such as Zamfirescu-Pereira et al. (2023) recommend, might increase the LLM reliability. Although such manipulation falls outside the scope of this exploratory study, it is a commendable direction for future research. Another limitation to build upon is investigating the effect on reliability of including longer text excerpts, perhaps section after section from full articles, in the prompt.

Looking ahead, it is likely that academic literature review will increasingly be conducted with the aid of LLMs and other AI approaches, such as machine learning and natural language processing. Davarathne et al. (2024) demonstrate how such technologies can be applied to enhance academic library search. Van de Schoot et al. present an 'active learning' algorithm that assists with sorting and selecting articles (2021, p. 125). It is even speculated that LLMs could be employed in the more creativity- and reasoning-dependent tasks of information synthesis, critical assessment, and argument development (Fok & Weld, 2023). Moreover, it is probable that novice and even experienced scholars wishing to expedite information synthesis and argument development will try to leverage GenAI tools to these ends. The need for more empirical research into the effectiveness and reliability of these technological applications, beyond mere proof-of-concept demonstrations, is of paramount importance at this juncture, to ensure that scholars are able to make good use of such tools. Such research will have the most benefit for novice scholars with unrefined judgement of the quality of automated computer program outputs.

At this point, our findings problematize the adoption of LLMs as a direct delegative scaffold for novices learning to conduct a literature review. However, that is not to say that GenAI-supported literature review strategies could not be conceived and tailored in the realm of academic writing pedagogy. Because AI-powered tools increasingly integrate LLMs as part of processing scholarly texts, it is critical for novices to develop the skills needed to both use and supervise these technologies, despite their limitations. Consequently, future research should inform writing teachers how to best devise a ‘human-in-the-loop’ approach in the literature review process. Rather than obviating the need to advance competency in reading comprehension, LLMs introduce the novel exigency of applying reading comprehension skills to the task of assessing AI-generated outputs. Scholarship pedagogy should emphasize reading skill development and help novice literature reviewers develop the ability to oversee LLM affordances, critically evaluate LLM outputs, and make meaning from the literature for themselves.

References

Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A. E. 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

Álvarez-Martínez, F. J., Borrás-Rocher, F., Micol, V., & Barrajón-Catalán, E. (2023). Artificial intelligence applied to improve scientific reviews: The antibacterial activity of *Cistus* plants as proof of concept. *Antibiotics*, *12*(2), 327. https://doi.org/10.3390/antibiotics12020327

Bedington, A., Halcomb, E. F., McKee, H. A., Sargent, T., & Smith, A. (2024). Writing with generative AI and human-machine teaming: Insights and recommendations from faculty and students. *Computers and Composition, 71*(2024), 102833. https://doi.org/10.1016/j.compcom.2024.102833

Bernhardt, L. (2023). Literature review basics. *David L. Rice Library Research Guides*. University of Southern Indiana. https://usi.libguides.com/literature-review-basics/tables

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, *9*(1), 2–20. https://doi.org/10.1016/j.jeap.2010.01.001

Boell, S. K., & Cecez-Kecmanovic, D. (2014). A hermeneutic approach for conducting literature reviews and literature searches. *Communications of the Association for Information Systems, 34*. Article 12. https://doi.org/10.17705/1CAIS.03412

Carver, J. C., Hassler, E., Hernandes, E., & Kraft, N. A. (2013). Identifying barriers to the systematic literature review process. *2013 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*. https://doi.org/10.1109/ESEM.2013.28

Chen, D.-T. “Victor”, Wang, Y-M., & Lee, W. C. (2016). Challenges confronting beginning researchers in conducting literature reviews. *Studies in Continuing Education*, *38*(1), 47–60. https://doi.org/10.1080/0158037X.2015.1030335

Chen, P., Poeppel, D., & Zuanazzi, A. (2023). Meaning creation in novel noun-noun compounds: Humans and language models. *Language, Cognition and Neuroscience, 39*(2). https://doi.org/10.1080/23273798.2023.2254865

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Consensus. (2024). AI Search Engine for Research. https://web.archive.org/web/20240711043247/https://consensus.app/. Accessed on July 12, 2024.

Davarathne, R. N., Arachchi, N. H., Aluthdeniya, N. D., Ranasinghe, W., & Ganegoda, G. U. (2024). Centralized and labeled academic journal library using machine learning & deep learning approaches. *International Conference on Image Processing and Robotics (ICIPRoB),* Colombo, Sri Lanka, 2024. https://doi.org/10.1109/ICIPRoB62548.2024.10544300

Elicit. (2024). Analyze research papers at superhuman speed. https://web.archive.org/web/20240711161101/https://elicit.com/. Accessed on July 12, 2024.

Fok, R., & Weld, D. S. (2023, April 23). *What can't large language models do? The future of AI-assisted academic writing* [Accepted paper]. In2Writing - The Second Workshop on Intelligent and Interactive Writing Assistants, Hamburg, Germany. https://cdn.glitch.global/d058c114-3406-43be-8a3c-d3afff35eda2/paper4_2023.pdf

Hoffman, M. J., Remus, S., Biemann, C., Radach, R., & Kuchinke, L. (2022). Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence, 4*, 730570. https://doi.org/10.3389/frai.2021.730570

Hope, T., Downey, D., Weld, D. S., Etzioni, O., & Horvitz, E. (2023). A computational inflection for scientific discovery. *Communications of the ACM, 66*(8), 62–73. https://doi.org/10.1145/3576896

Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology, 15*(4), ep464. https://doi.org/10.30935/cedtech/13605

Kacena, M. A., Plotkin, L. I., & Fehrenbacher, J. C. (2024). The use of artificial intelligence in writing scientific review articles. *Current Osteoporosis Reports*, *22*(1), 115–121. https://doi.org/10.1007/s11914-023-00852-0

Khalifa, M., & Albadawy, M. (2024). Using artificial intelligence in academic writing and research: An essential productivity tool. *Computer Methods and Programs in Biomedicine Update, 5*(2024), 100145. https://doi.org/10.1016/j.cmpbup.2024.100145

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, Y.-S. G. (2020). Interactive dynamic literacy model: An integrative theoretical framework for reading and writing relations. In R. Alves, T. Limpo, & M. Joshi (Eds.), *Reading-Writing Connections: Toward Integrative Literacy Science* (pp. 11-34). Springer. https://doi.org/10.1007/978-3-030-38811-9_2

Knowles, A. M. (2024). Machine-in-the-loop writing: Optimizing the rhetorical load. *Computers and Composition, 71*(2024), 102826. https://doi.org/10.1016/j.compcom.2024.102826

McHugh, M. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Melnyk, B. M., Gallagher-Ford, L., & Fineout-Overholt, E. (2016). *Implementing the Evidence-Based Practice (EPB) Competencies in Healthcare: A Practical Guide for Improving Quality, Safety, and Outcomes*. Sigma Theta Tau International.

Miao, Y., Lin, G., Hu, Y., & Miao, C. (2019). *Reading comprehension ability test-A Turing test for reading comprehension*. ArXiv. https://doi.org/10.48550/arXiv.1909.02399

Mollick, E. (April 26, 2023). A guide to prompting AI (for what it is worth). *One Useful Thing*. Substack. https://www.oneusefulthing.org/p/a-guide-to-prompting-ai-for-what

Ngwenyama, O. & Rowe, F. (2024). Should we collaborate with AI to conduct literature reviews? Changing epistemic values in a flattening world. *Journal of the Association for Information Systems, 25*(1), 122–136. https://doi.org/10.17705/1jais.00869

OpenAI. (2023a). *GPT-4 Technical Report*. ArXiv. https://doi.org/10.48550/arXiv.2303.08774

OpenAI. (2023b). *GPT-3.5 Turbo fine-tuning and API updates*. https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/

OpenAI (2024, May 13). *GPT-4o*. https://openai.com/index/hello-gpt-4o/

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinand's, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence, 3*(2021), 125–133. https://doi.org/10.1038/s42256-020-00287-7

SciSpace (2024). Do hours worth of reading in minutes. https://web.archive.org/web/20240707225800/https://typeset.io/. Accessed on July 12, 2024.

Scite (2024). AI for research. https://web.archive.org/web/20241001041904/https://scite.ai/. Accessed on October 2, 2024.

Susnjak, T., Hwang, P., Reyes, N. H., Barczak, A. L. C., McIntosh, T. R., & Ranathunga, S. (2024). *Automating research synthesis with domain-specific large language model fine-tuning*. ArXiv. https://doi.org/10.48550/arXiv.2404.08680

Tang, B. (2020). The Chiron imperative – A framework of six human-in-the-loop paradigms to create wise and just AI-human centaurs. In S.A. Bhatti, S. Chishti, A. Datoo and D. Indjic (Eds.) *The LegalTech Book*. Wiley. https://doi.org/10.1002/9781119708063.ch10

Toner, H. (2023, May 12). What are generative AI, large language models, and foundation models? *Center for Security and Emerging Technology, Georgetown University*. https://web.archive.org/web/20240913225524/https://cset.georgetown.edu/article/what-are-generative-ai-large-language-models-and-foundation-models/

Wagner, G., Lukyanenki, R., & Paré, G. (2022). Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology, 37*(2), 209–226. https://doi.org/10.1177/02683962211048201

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *The Journal of Child Psychology and Psychiatry, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Zamfirescu-Pereira, J. D., Wong, R., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23-28, 2023, Hamburg, Germany*. https://doi.org/10.1145/3544548.3581388

Appendix A: Large language model prompt

You are a social science scholar. You are researching the impacts that artificial intelligence innovations have had on academic writing as preparation to write a journal article. You are conducting a literature review, and you are using a principled method to understand the content of a peer-reviewed journal article. Here is the first article:

[Text from the first page of each journal article]

For this article, you are trying to identify the purpose of the article, what type of article it is (theory, research, etc.), and what paradigm it adopts (quantitative, qualitative, mixed methods). For empirical research articles, you also classify the research design and the major findings. Finally, you will assign each paper a “Level of Evidence” from 1-7, based on the research design, as follows: 1=systematic reviews of randomized controlled trials (RCTs); 2=randomized controlled trials (RCTs); 3=controlled trials without randomization; 4=case-control and cohort studies; 5=systematic reviews of descriptive and qualitative studies; 6=single descriptive or qualitative studies; 7=expert opinions. If the research follows a clinical trial model, identify the population of interest, intervention or issue of interest, comparison intervention or group, outcome (including statistical significance), and timeframe, following the PICOT model. Present your results in the format of a python dictionary, as follows: {authorYear: [article author and publication year], studyPurpose: [purpose of study], researchQuestions: [research questions], articleType: [type of article], paradigm: [paradigm], design: [design], LOE: [level of evidence], findings: [major findings], population: [population of interest], intervention: [intervention], comparison: [control group treatment], outcome: [outcome], timeframe: [timeframe]}. If any of these values are not in the provided article, leave them empty. Do not output anything other than the python dictionary. Be sure to enclose the property names, i.e., dictionary keys, in double quotes.

Appendix B: Example LLM Output

Content Category	GPT-3.5 Turbo Output	GPT-4o Output
authorYear	['Chauke et al.', 2024]	['Chauke et al., 2024']
studyPurpose	explore postgraduate students' perceptions of the benefits associated with the utilisation of artificial intelligence tools, with a specific focus on ChatGPT, in their academic success at historically disadvantaged universities in South Africa	["To explore postgraduate students' perceptions of the benefits associated with the utilisation of artificial intelligence tools, specifically ChatGPT, for their academic success at South Africa,Äôs historically disadvantaged universities."]
articleType	research	['Research']
paradigm	qualitative	['Qualitative']
Research Design	qualitative study	['Semi-structured interviews with thematic analysis']
LOE	6	[6]
findings	significant advantages of incorporating ChatGPT in students' academic journey, particularly in research success, helping with refining research topics, identifying grammatical errors, and enhancing academic writing skills	['ChatGPT helps refine research topics, identifies grammatical errors, and assists in paraphrasing academic writing, thus enhancing academic writing skills.']
population		[ ]
intervention		[ ]
comparison		[ ]
outcome		[ ]
timeframe		[ ]