Can You Trust Artificial Intelligence to Conduct a Literature Review in Accounting Research?

Derek K. Oler*
derek.oler@ttu.edu

Shengbai Zhang*
shengzha@ttu.edu

Yan Zhangβ
yz@nmsu.edu

Ye Zhang*
zha38644@ttu.edu

We investigate and compare the accuracy of various AI models in conducting a literature review. We begin with ChatGPT (including ChatGPT Plus) and Gemini, and compare their results with results from a more recent model, Perplexity, in conducting a literature review using 33 accounting research prompts and verifying the accuracy of papers cited. ChatGPT cited 334 papers, but 31 citations (9.3%) had errors and 59 (17.7%) were hallucinations. ChatGPT Plus cited the fewest papers, 221, with 80 citation errors (36.2%) and 42 (19.0%) hallucinations. Gemini cited the most papers (405) but had the most citation errors (31.9%) and the most hallucinations (22.2%). Perplexity cited 269 papers with 70 errors (26.0%) and fewest hallucinations (3.3%). Accuracy improves across all models at the cost of longer time needed and subscription fees when we use "Deep Research." Accuracy is best when we use an even newer model, AIRA.

October 30, 2025

* Lyons School of Accounting, Rawls College of Business, Texas Tech University. We thank Gary Fleischman, Chris Skousen, Kirsten Cook, David Wood, Hamid Vakilzadeh, and various AI models for on comments and suggestions on earlier versions of this paper. All errors are the responsibility of the human authors.
β Accounting and Information Systems, New Mexico State University.

Can You Trust Artificial Intelligence to Conduct a Literature Review in Accounting Research?

1. Introduction

Generative and large-language AI models have gathered immense attention from academics in recent years. We test the accuracy of multiple artificial intelligence (AI) models in conducting a literature review on 33 topics in accounting research: ChatGPT (both the free version and subscription-based ChatGPT Plus), Gemini, and a more recent agentic AI model Perplexity, and we compare our results across models with those from the latest AI model (at the time of writing): "Artificial Intelligence Research Assistant" (AIRA). Both Perplexity and AIRA began as RAG models ("Retrieval-Augmented Generation" models)[1] and have evolved into agentic AI models.[2]

The term "generative" refers to the function of AI models, because generative models create new text and sometimes other content, such as images, video, and audio files, based on user prompts. "Large-language" (LLM) refers to the core architecture of the model, where "large" refers to the vast amounts of data the model draws from and "language" refers to the model communicating in human-like language. All models we investigate are both generative models and LLMs.

[1] Retrieval-augmented generation is a technique that enables large language models to retrieve and incorporate new information. With RAG, large-language models (LLMs) respond to user queries by referencing a specified set of documents. These documents supplement information from the LLM's pre-existing training data; RAG improves LLMs by incorporating information retrieval before generating responses. See https://en.wikipedia.org/wiki/Retrieval-augmented_generation

[2] Agentic AI systems (or models) are a class of artificial intelligence that focuses on autonomous systems that can make decisions and perform tasks with limited or no human intervention. Agentic AI is sometimes referred to as an "AI agent." See https://en.wikipedia.org/wiki/Agentic_AI

We take a user's perspective in evaluating these AI-powered models even though they are structured differently. This paper investigates how well ChatGPT, Gemini, Perplexity, and AIRA perform in performing the unstructured task of listing prior research on a particular accounting research topic. Our investigation is relevant to accounting research faculty and PhD students who consider starting a literature review on a possible research topic by prompting an AI model (sometimes multiple models) to list prior relevant papers in that area, or perhaps even drafting the text of a literature review for a paper. For example, a researcher might query ChatGPT on prior accounting research papers on the accruals anomaly, or on auditor turnover.

However, generative AI can make errors and even hallucinate. A hallucination is when AI claims something that does not exist; in our setting, a research paper citation that does not correspond to any published paper or working paper (we call these "fakes" or "hallucinations"). Prior accounting-focused work has evaluated ChatGPT's performance in answering accounting questions at both the undergraduate and masters degree level (Wood et al., 2023).[5]

We entered a total of 33 prompts into ChatGPT in late June / early July 2025, using the free version of ChatGPT 4o (4 "Omni").[6] We also interrogate Gemini (Gemini 2.5 "flash," the free version offered by Google) and Perplexity (free version) using the same prompts, over July and early August 2025. These prompts reflect typical accounting research topics (e.g., the accruals anomaly, financial statement credibility, income taxes, and auditor turnover). Responses per

[5] At the time of this paper's data collection the version available was ChatGPT 4o (4 "Omni"); at the time of Wood's (2023) paper data collection the version available was ChatGPT 3.5.

[6] A newer version, ChatGPT 5, was launched on August 7, 2025, just after our ChatGPT data collection ended. In all cases we used the most up-to-date AI model available the date of collection, and the more recent update of ChatGPT underscores the quickly evolving nature of generative AI.

prompt varied from 5 (Perplexity) to 30 (Gemini) papers. ChatGPT produced an average of 10.1 papers cited per prompt and a total of 334 papers, compared with 6.7 average per prompt (221 total) for ChatGPT Plus, 12.3 average (405 total) for Gemini, and 8.2 average (269) for Perplexity.

For ChatGPT, 73.1% of these citations were completely correct (244 papers citations out of 334), meaning that we could independently verify the paper, but 3.9% (13), 5.1% (17), and 0.3% (1) had errors in the citation (incorrect or missing journal name, year/volume/issue, or a misleading characterization of the paper's topic and contribution, respectively). Another 59 of these citations (17.7%) were fakes or hallucinations where cited paper did not exist.

ChatGPT Plus cited the fewest papers, 221, and 44.8% of these citations were completely correct (99 papers citations out of 221); 14.0% (31), 16.3% (36), and 5.9% (13) had errors in the citation (incorrect or missing journal name, year/volume/issue, or a misleading characterization of the paper's topic and contribution, respectively). Another 42 of these citations (19.0%) were hallucinations.

For Gemini, 45.9% of the papers cited were completely correct (186 citations), 31.9% had errors in the citation (129), and 22.2% were fake (90). For Perplexity, 70.6% were completely correct (190), 26.0% had errors (70) and only 3.3% (9) were hallucinations.

There are additional research options available for some AI models; specifically, we selected the "Deep Research" and "Think Longer" options available for ChatGPT (AI uses the term "tools")

from a dropdown menu available at the OpenAI website. The Think Longer tool in ChatGPT is a user interface feature enabling the model to generate more thoughtful, detailed, and refined responses. Our use of Think Longer for some of the same prompts takes slightly longer (a few more seconds) but resulted in fewer hallucinations (i.e., citations of fake papers) and fewer errors in citations for real papers. The Deep Research tool is available for ChatGPT, Gemini, and Perplexity and is a premium feature (offered for free for a few first uses) that enables the model to gather high-quality, up-to-date information from the web to answer complex or niche questions more thoroughly. Our use of Deep Research for some of the same prompts takes considerably longer but results in higher accuracy in results. Both think longer and Deep Research invoke agentic AI. At the time of writing OpenAI limits free account users to a total of 5 deep searches. Gemini and Perplexity also offer Deep Research for free for the first few prompts, and we find that Deep Research results in far fewer errors and hallucinations for these models as well.

Finally, we also interrogated AIRA ("Artificial Intelligence Research Assistant," a newer AI model) using the same four prompts. AIRA was able to produce an answer slightly faster than Perplexity's Deep Research tool and cited a total of 60 papers (an average of 15 per prompt). Overall accuracy was highest (86.7% error-free citations), with 8 errors (duplicate or misleading citations), and no hallucinations.

2. Background and Prior Literature

ChatGPT is produced by OpenAI, and OpenAI was founded in December 2015 by Elon Musk and other collaborators to ensure that artificial general intelligence (highly autonomous systems that outperform humans at most economically valuable work) benefits all of humanity. [7] ChatGPT was first available to online users on November 30, 2022. Gemini is a Google AI model that first became available on December 6, 2023 (first named Bard and later renamed Gemini), and Perplexity is an AI model produced by Perplexity AI Inc that was first available in December 2022. [8] Perplexity started as a RAG (retrieval-augmented generation) system, meaning that Perplexity draws data from a number of large-language models (LLMs), described in more detail below, by first searching its own indexed data for information relevant to the user's prompt, retrieving that indexed data, and then assembling English responses for the user. Unlike other generative AI systems that respond to prompts based on the data they have "learned" (i.e., data entered when the model was trained), RAG systems begin by retrieving data from an external database and next generate a response to the user's question or prompt. We use the "base" model of Perplexity that draws from ChatGPT, Claude, Gemini, Grok, and Perplexity's own LLM named Sonar.[9] RAG systems are less likely to produce hallucinations but are not necessarily error-free. AIRA was created by Hamid Vakilzadeh at the University of Wisconsin-Whitewater and David Wood at BYU to assist researchers in navigating and synthesizing the growing number of research articles in a given area and was first available for public use in September 2023. They provide a thorough description of their work and RAG system technology in Vakilzadeh and Wood (2025), forthcoming in the *Journal of Information*

[7] Quote from query "why was open AI founded?" from ChatGPT 4o. We use AI quotes and statements throughout this paper. ChatGPT can be accessed at https://chatgpt.com/.
[8] Gemini can be accessed at www.google.com (select from the "google apps" dropdown menu at the top left hand corner), and Perplexity can be accessed at www.perplexity.ai).
[9] The "Pro" model offers user options to select the different LLMs that Perplexity draws from.

Systems. AIRA is housed by Semantic Scholar, which is part of Allen Institute for AI. AIRA also started as a RAG and, like Perplexity and has evolved into an agentic agent.[10] Perplexity and AIRA are different in that Perplexity is an AI-powered "answer engine" that combines a web search with natural language responses (some would describe Perplexity as "super Google") while AIRA is a research-focused AI agent that specializes in summarization and information synthesis. AIRA draws data primarily from the Semantic Scholar database.

Agentic AI refers to autonomous AI systems (or "agents") that can act independently to achieve pre-determined user-specified goals. These independent actions include invoking tools or executing program code, and they can even coordinate with other AI agents (Wu et al, 2024). Traditional artificial intelligence requires prompting and step-by-step guidance from the user; agentic AI is proactive and can perform complex tasks without constant human oversight.[11]

ChatGPT and Gemini are part of a family of generative artificial intelligence models. Perplexity and AIRA are more advanced generative AI models that started as RAG models but both have evolved into agentic AI models. The term "generative" refers to the function of the models, that is, to create or generate new material in response to user prompts. These AIs are also chatbots, or large language models (LLMs), and are trained to predict the next word in response to a query, allowing it to provide English-sentence-answers. However, especially for ChatGPT and Gemini, this training methodology can also give rise to hallucinations, or in our case, citing papers that

[10] At the time of writing access to AIRA was available at https://smithery.ai/server/@hamid-vakilzadeh/mcpsemanticscholar. From that web page users can enter prompts after first clicking on "explore capabilities" in the top right-hand corner of the page.
[11] See https://aws.amazon.com/what-is/agentic-ai/

don’t exist, because there is no automatic checking in these models to validate the response given with outside information. Hallucinations are false or misleading information that sound plausible. In contrast, the advantage of RAG models is that they *do* verify data they obtain from AI models with external databases. This does not guarantee error-free responses, but should significantly reduce errors and hallucinations.

One anecdotal case illustrates the problems of AI use in accounting research. A paper recently submitted to a top accounting journal had 1/3 of its cited papers that were hallucinations; this was discovered when a reviewer saw a citation for a paper by two people he knew personally, but he had not heard of them collaborating on a paper. When he checked, the cited paper did not exist. This prompted checking *all* the papers cited, and the discovery of 1/3 of them being fake.

ChatGPT lists several reasons for why it may produce hallucinations (and the same explanations apply to the other models we use):[12]

1. It's a pattern predictor, not a fact checker

- ChatGPT is trained to predict the next word in a sentence based on patterns in vast amounts of text.
- It doesn't know facts like a human and it doesn’t check by default (unless tools like web search are used).
- It might confidently state something that looks statistically likely, even if it's completely wrong.

2. Training data isn't always accurate

- The model learns from internet-scale text, and this includes errors, outdated information, and conflicting sources.

[12] This is another response from ChatGPT to the prompt “why can ChatGPT produce hallucinations?”

- If enough texts mention something incorrectly, the model may learn and reproduce that mistake.

3. It has no built-in memory

- ChatGPT doesn't "remember" where it read something. It generates answers based on patterns.
- This can lead to made-up quotes, fake references, or false statistics that sound real.

4. Pressure to Respond Confidently

- The model is designed to give answers even when uncertain (PhD programs often have the same effect on students!).
- If asked something obscure, it may "fill in the blanks" rather than admit it doesn't know, and this can cause hallucinations.

5. Lack of Real-Time Verification

- Unless tools like web search or Deep Research are used, ChatGPT operates on a static snapshot of knowledge.
- It doesn't have live access to databases or APIs (Application Programming Interfaces, a set of rules and tools that lets different software systems communicate with each other) to double-check facts.

6. Prompt Ambiguity

- Vague or leading questions can cause the model to "guess" what the user wants to hear, sometimes creating inaccurate or fabricated content to match expectations.

Gemini is focused more on generating human-like text and is more loquacious than ChatGPT, Perplexity, and AIRA. Perplexity is described as an "answer engine" designed to combine up-to-date web searches with natural language answers. AIRA is a research-focused program ("agent") used for analyzing, summarizing, and synthesizing academic papers. Perplexity and AIRA originated as RAGs ("Retrieval-Augmented Generation") systems that can draw real-time data from external databases and are much less likely to hallucinate.

Wood et al (2023) report that ChatGPT performed relatively well in answering true/false and multiple-choice questions (its accuracy rates were 68.7% and 59.5%, respectively, a little worse than human scores), when crowd-sourced coauthors entered their exam questions into ChatGPT and reported the accuracy of its responses. Dong, Stratopoulos and Wang (2024) review ChatGPT research in accounting and finance but do not look at using ChatGPT for literature reviews. They advise: "the authors caution accounting professionals against an over-reliance on these new technological tools, warning that such over-dependence could potentially undermine their critical thinking skills…[We] recommend that CPAs exercise professional skepticism and use their LLMs to enhance rather than replace their expertise." (pages 10-11).

Somasundaram (2023) and Stapleton (2024) provide a general step-by-step guide to using ChatGPT to write a literature review, and Somasundaram advises "…it's important to critically evaluate and validate the information it generates."

Haman and Skolnik (2024) discuss the use of ChatGPT for literature reviews for medical research and conclude "…we firmly recommend not using ChatGPT in the research process." They report the results of using 5 unique "chats" with the prompt "List 10 seminal academic articles in the field of medicine and provide the DOI."[13] Each chat resulted in ChatGPT providing 10 publications for a total of 48 unique publications; that is, ChatGPT was not consistent in its replies to exactly the same prompt over time. When they checked each paper citated they found that 66% of those papers did not exist (i.e., they were hallucinations). Even for the real papers, ChatGPT sometimes gave an incorrect DOI. In discussing the ethics of using

[13] DOI stands for Digital Object Identifier, a unique string of characters used to permanently identify an object, such as a journal article or research data, and provide a stable link to its location on the internet.

ChatGPT for a literature review, Haman and Skolnik (2024) note that the authors of a publication must be held responsible for that review, not ChatGPT.

Wood (2025) notes in his insightful book, *Rewiring Your Mind for AI,* one of many well-publicized cases where lawyers used ChatGPT to draft legal briefs without realizing (or checking) that the AI was citing fictitious precedents (see Neumeister 2023, Skolnik 2024, and Wolf 2025 for examples of legal problems from inappropriate ChatGPT use by lawyers). In the case cited by Wood, ChatGPT fabricated six cases that appeared completely real. Interestingly, the lawyer also asked ChatGPT if it was fabricating the cases it was citing, and ChatGPT responded that it was not! AI has no internal mechanism that can distinguish between reality and fantasy.

3. Methodology

We test the listing of papers provided by ChatGPT, Gemini, and Perplexity in response to 33 accounting research topic prompts as an objective means to evaluate each model's accuracy. Responses to literature review prompts can also be evaluated based on their comprehensiveness (do they include the best, most relevant papers?) and appropriateness (do they draw correct conclusions on main findings of each paper?), but those evaluations are more subject to nuance and reader discretion. In contrast, it is relatively easy to determine if a cited paper exists or not, and if the citation provided is correct.

We recorded a paper as real even if the journal name, year, volume or issue was incorrect, so long as the title and authors of the paper were correct. However, if we could not locate the cited paper (even in a different journal), we recorded the citation as a fake or hallucination.

We performed the following steps to gather data to evaluate each model:

1. We entered the prompt into the model (the prompt list is given in the Appendix). For ChatGPT we used the following URL: https://chat.openai.com. For Gemini we went to www.google.com and selected Gemini from the dropdown option list on the top right. For Perplexity we went to https://www.perplexity.ai/.

   At the time of data collection (late June to early August of 2025) the most advanced but free ChatGPT model available was ChatGPT 4o ("omni"), although users must sign up for a free account.[14] The most recent version of Gemini was Gemini 2.5 Flash. Perplexity does not use traditional version numbers, and when questioned Perplexity states that it continuously updates its platform and features without assigning a public version number.

2. For each paper cited in the response, we verified the citation using our school's library web page to search online journals.
   - We searched for the journal, year, volume, and issue given in the citation. If we located the paper, we marked the citation as "real."

3. If unable to find that paper in the above step, we next did a Google search for the paper's title. This would sometimes produce the correct paper, often in a different journal, year, volume, or issue, and more rarely as an unpublished working paper.

4. If we located the paper (that is, the same title and authors) in a different journal for a different year, volume, or issue, we noted that the citation was incorrect but the paper was real.

5. If we could not locate the paper in the above steps we would search for the CV of one of the authors and try to locate the paper (or a similar paper) in her CV. If the paper was listed in her CV we noted that the paper was real.

6. If we could not locate the paper after the above steps we recorded that the paper citation was fake.

[14] Strangely, sometimes when we asked ChatGPT which version we were chatting with, the answer would vary. At times it would claim to be ChatGPT 3.5, and at other times it would claim to be ChatGPT4o. A newer version of ChatGPT GPT-5, came out on August 7, 2025.

We report our main results in Table 1. We next checked if our results can be improved by using the Think Longer (available for ChatGPT) and Deep Research options (available for ChatGPT, Gemini, and Perplexity). Because these were more time consuming queries we limited ourselves to checking four different prompts. We compared the responses to our regular query used in Table 1, with results for the Think Longer and Deep Research tools in Table 2. Gemini and Perplexity do not offer the Think Longer tool but do have Deep Research tools, and we report the results for those tools in Tables 3 and 4. Finally, AIRA is a relative newcomer to the AI tools available to academics, and we show results for using AIRA for the same four prompts, alongside results for ChatGPT, Gemini, and Perplexity, in Table 5.

4. **Results**

We find that AI model output is not consistent, similar to the findings of Haman and Skolnik (2024). That is, it is not possible to reproduce the same output from the AI model even with exactly the same prompt, even if that prompt was entered a short time later by the same user. This reflects that generative AI models are constantly evolving because the underlying software updates itself based on a constant stream of new inputs.

Our Appendix lists the prompts we used. Our listing reflects an evolution in prompts where the first two prompts were vague ("accounting research on CEO and CFO turnover" and "business research on CEO and CFO turnover"), although they both produced lists of paper citations that we could verify. Our next prompt was more specific ("List prior accounting research papers on

the accruals anomaly") but the response did not contain complete citations and required more time to verify, so we added "with citations" and later added "and URLs" in our later prompts. These did not prevent hallucinations or mistakes in the citations but allowed us to verify the responses more efficiently.

[Insert Table 1 Here]

We entered the 33 prompts listed in our Appendix into each model at various dates from late June to early August 2025. Responses from ChatGPT varied from citing 7 to 16 papers, with an average of 10.1 papers and a grand total of 334 papers. Responses from ChatGPT Plus varied from citing 4 to 10 papers, with an average of 6.7 papers and a total of 221 papers. Responses from Gemini ranged from 5 to 30 papers, averaging 12.3 papers per prompt and a total of 405 papers. Responses from Perplexity varied from 5 to 13 papers and a total of 269 papers.

We followed the above steps for each citation for each model and report the results in Table 1. One example of an error in the citation for a "real" paper was when ChatGPT listed a citation as 2006, *JAE* 40(1-3), but the correct citation was 2006, *JAR* 44(2).[15] ChatGPT reported another citation as *JAR* 2007 45(2) but should have been cited as *JAR* 2006 44(2).[16] A final example of a mistake from ChatGPT was when a paper was listed as examining the relationship between CEO turnover and firm performance, but the actual paper examines group affiliations in diversified Indian groups.[17]

[15] The correct paper is "Accounting discretion in fair value estimates: An examination of SFAS 142 goodwill impairments" written by A. Beatty and J. Weber, *Journal of Accounting Research* 2006 volume 44 issue 2.
[16] The paper is Kraft, Leone, and Wasley (2006)
[17] See Khanna and Palepu (2000).

Most significantly for ChatGPT, 59 of the 334 papers (17.7%) were hallucinations. For example, in response to the prompt "list prior accounting research papers on fraud with citations," ChatGPT cited a fake paper titled "Detection of financial statement fraud and other forms of corporate misconduct: A review of the literature" by Petrols, Bowen, Zimmerman, and Peng in *Auditing: A Journal of Practice & Theory* 2006 36(2). The closest real paper seems to be "Financial Reporting Fraud and Other Forms of Misconduct: A Multidisciplinary Review of the Literature" by Amiram, Bosanic, Cox, Karpoff, and Sloan published in *Review of Accounting Studies,* 2018 23(2).

ChatGPT Plus, a subscription-based AI model, cited the fewest papers in response to the same 33 prompts (221). Proportionately more of these were hallucinations (19% or 42 papers), and 36.2% had errors (179). Only 99 of the citations (44.8%) were completely correct.

Gemini appears to be the most conversational and verbose of the models we used and provided the most citations per prompt (and the greatest variation in the number of citations), but also has (at the time of data collection) the highest error rate. For example, in response to the prompt "list prior accounting research papers on the artificial intelligence with citations and URLs" Gemini produced a list of 12 papers, 11 of which were hallucinations, and the one non-hallucination still had the wrong year, volume, and page numbers (but the paper title and authors were correct). One of the hallucinations was a fake paper titled "An overview of machine learning in accounting research" by Gleim and Gomaa, in *The Accounting Review* 2020, 95(1), pages 1-28. We could not find any remotely similar real paper.

Perplexity was the most accurate of the models we used (at the time of data collection).[18] For example, in response to the same AI prompt above ("list prior accounting research papers on the artificial intelligence with citations and URLs"), Perplexity produced a list of 13 papers that were all real, but 2 citations were missing author and journal names.

To see if relatively small adjustments can improve the accuracy of ChatGPT's responses we also used the options "Think Longer" and "Deep Research." Recall that both of these options invoke agentic AI where AI "agents" or autonomous programs operate more independently in response to user requests. We selected these options from the "tools" dropdown bar at the website https://chatgpt.com/. We selected four prompts to revisit using these more rigorous options, with results shown in Table 2.[19] We repeated the same process for Gemini and Perplexity. Gemini and Perplexity do not offer the Think Longer tool, but offer the Deep Research tool. Results using these tools are shown, alongside results for prompts with no options selected for comparison, are shown in Tables 3 (Gemini) and 4 (Perplexity).

[Insert Table 2 Here]

"Think Longer" instructs the model to spend more time reasoning, consider more angles or complexities, expand and elaborate on answers, and (hopefully) improve the quality and depth of

[18] Although we had lower error rates later when we used AIRA, at the cost of fewer papers cited.
[19] The prompts were "list prior accounting research papers on the causes and effect of CEO and CFO turnover on firm performance with citations and URLs," "list prior accounting research papers on taxes with citations and URLs," "list prior accounting research papers on fair value accounting with citations and URLs," and "List prior accounting research papers on tax avoidance with citations and URLs."

the response. To facilitate comparisons, Table 2 Panel A reports the results to our no option search (i.e., the same as reported in Table 1, Panel A, limiting our search to only the four prompts listed earlier). Panel B shows our results for the "Think Longer" tool and Panel C shows results for the "Deep Research" tool. Our original queries produced responses of 39 papers, for an average of 9.8 citations per prompt. Out of these, 26 citations were correct (66.7%), 6 were hallucinations (15.4%), and 7 had errors (17.9%). Results from the Think Longer tool were better (Panel B), but only 36 papers were cited. Out of these, 28 (77.8%) were completely correct, only 2 (5.6%) were hallucinations, and 6 (16.7%) had other errors. Deep Research produced the most citations (44), and out of these, 36 (81.8%) were completely correct, 8 (18.2%) had errors (wrong journal name, etc.), and none were hallucinations.

We checked Gemini's Deep Research tool for the same four prompts and show our results in Table 3. Our original no option search, reproduced in Table 3 Panel A for comparison, produced 46 citations, 23 of which (50%) were completely correct and 19 (41.3%) had minor errors, and 4 were hallucinations (8.7%). The Deep Research option produced more paper citations (105, Panel B), and had relatively fewer errors (26 or 24.8%). Most of these errors were because Gemini did not include URLs or complete citations in its response, even though explicitly prompted to do so. Our Deep Research results had fewer hallucinations (2 or 1.9%).

[Insert Table 3 Here]

Our "no option" results for Perplexity were the most accurate of our no option results: Table 1 Panel D reports that 70.6% of the papers cited were completely correct, and only 3.3% were

hallucinations. But we also checked Perplexity's results using the Deep Research tool, in Table 4.

[Insert Table 4 Here]

The same four prompts produced many more citations when we used the Deep Research tool (104 instead of 33). Slightly fewer were completely correct (76.6% or 80), and another 23 (22.1%) had errors, but only 1 citation (1%) was a hallucination. This is not surprising given that Perplexity is a RAG model that offers the additional benefits of drawing data from multiple AI models and using objective verification of the results from each model.

Finally, we examined the performance of AIRA using the same four prompts. Recall that AIRA is another RAG model developed by two accounting professors (Hamid Vakilzadeh and David Wood), and the details of AIRA are given in Vakilzadeh and Wood (2025); in this paper we compare results from AIRA with the other three models we examine. To facilitate this comparison we repeat the results from our Deep Research tool for ChatGPT (Panel A), Gemini (Panel B), and Perplexity (Panel C).[21] AIRA does not offer a Deep Research tool. Our results are shown in Table 5.

[Insert Table 5 Here]

[21] ChatGPT Plus did not perform as well as the base ChatGPT so we did not examine Deep Research results for ChatGPT Plus.

AIRA provided fewer papers but was relatively more accurate: 60 papers cited (average of 15 per prompt), but search results took less than a minute to produce. Out of the 60 citations 52 (86.7%) were error-free, 2 were duplicates, and 6 were misleading (the paper's topic was not consistent with the prompt). We did not find any hallucinations in the AIRA results.

5. Discussion and Conclusion

Overall, AIRA and Perplexity performed well in response to less structured prompts that simulate an accounting researcher's first steps in conducting a literature review. ChatGPT 4o (both the free model and the subscription-based "Plus" model) was next in terms of accuracy, and Gemini was the least accurate (but produced the greatest number of citations).[22] Our results suggest improvement in ChatGPT's performance relative to the findings from 2023 reported by Haman and Skolnik (2024); recall that they reported a 66% hallucination rate, as compared with our rate of "only" 17.7% (for the free version) and 19.0% for the "plus" version. Our observation that in many cases the citation was correct but the details (often the URL) had mistakes underscores that accounting research papers often begin as working papers (usually posted on SSRN) that later appear as published papers in a number of journals, sometimes with a slightly different title or even with a new coauthor added. Many of these iterations will appear on the internet, but AI models have difficulty identifying iterative versions of the same paper.

Our work provides a timely warning to accounting researchers: AI models are a convenient and helpful resource, but do not accept AI responses without first verifying their accuracy!

[22] We stress that these accuracy statistics are as of the time of data collection (summer 2025), because AI models constantly evolve and (hopefully) constantly improve.

Much of what we report here is likely already anecdotally known (or suspected) by many, but we offer objective documentation and guidance to researchers. Further, we make no effort to dive into the weeds of evaluating how appropriate a given prompt response is for the researcher's needs. We don't prompt our AI models for the "best" paper in a given area (presumably anyone who as published anything in a given area thinks his paper is the best). Our results suggest that multiple prompts and queries should be used to get a comprehensive list.

AI use is still controversial. A mainstream media article (Allbert, 2025) cited a recent MIT study on the effect of AI use (using the ChatGPT model) on students. The study compared EEG brain scans of students writing essays with and without using AI, and reported that AI users showed significantly lower neural engagement and had less ability to recall what they had written even moments earlier.[23] Anecdotally, others disagree with the study's findings and report that AI use seems to increase rather than decrease cognitive engagement and creativity. In another AI-related scandal involving accounting, Deloitte recently refunded $63,000 (US) to the Australian government after a report they provided was found to have multiple AI-generated errors, including a fabricated quote from a court judgement and references to fake academic papers (Alexis, 2025).[24]

We suggest the following steps to ensure a more credible and accurate response from AI models:

[23] Available at https://www.theepochtimes.com/health/the-cognitive-debt-were-accumulating-every-time-we-use-ai-5889854?utm_source=Health&src_src=Health&utm_campaign=health-2025-07-28&src_cmp=health-2025-07-28&utm_medium=email&est=HU3kfjFFEegkWYni1k5cQnYEojix42PdsqKMSxIqSeaWH2%2Baxs5y%2BxGv7w%3D%3D

[24] Available at https://www.cfodive.com/news/deloitte-refunds-60k-report-ai-errors-australian-government-accounting/803321/

1. When possible, use AIRA or Perplexity (as of the time of writing, 10/29/2025), although all AI models continue to rapidly evolve.
2. When time allows, use the "Deep Research" option (available in three of our AI models tested, although each model requires a subscription fee for entering more than a few prompts).
3. For convenience in verifying the papers in the response, include the phrase "include the URL" as part of the prompt. Note that this will not guarantee accurate results, because sometimes the link provided was to an unrelated website that does not make any mention of the paper. Gemini sometimes did not produce URLs even though explicitly prompted to do so.
4. Compare the results of multiple queries using similar prompts.[25]
5. Read each response carefully – DO NOT cut-and-paste an unvetted response into your paper!

Finally, researchers should consider the ethical implications of using AI in their research. At minimum, AI use should be acknowledged (as we do), and any AI response should be checked for plagiarism and errors. AI should not be used in circumstances where it is prohibited (for example, in producing a deliverable for a PhD seminar).

The authors are aware of another anecdotal case where the authors correctly disclosed AI use in drafting their paper. A journal editor asked about how specifically AI was used, and the authors responded that AI had written multiple drafts of the paper (that the authors subsequently read and suggested changes back to the AI). The AAA is currently wrestling with how AI should be used, and disclosure of its usage, in accounting research.

[25] We expect that other AI models will quickly become available, including the recently updated ChatGPT 5.0. Although not powered by AI, www.litmaps.com offers a pictographic literature review of related papers that researchers can use (a free version is available with limitations).

Works Cited

Allbert, Makai. 2025. The cognitive debt we accumulate every time we use AI. *The Epoch Times*, July 26, 2025. Available at https://www.theepochtimes.com/health/the-cognitive-debt-were-accumulating-every-time-we-use-ai-5889854?utm_source=Health&src_src=Health&utm_campaign=health-2025-07-28&src_cmp=health-2025-07-28&utm_medium=email&est=HU3kfjFFEegkWYni1k5cQnYEojix42PdsqKMSxIqSeaWH2%2Baxs5y%2BxGv7w%3D%3D.

Alexis, Alexei. 2025. Deloitte refunds over $60K for report with AI errors, Australian government says. *CFO Drive,* October 21, 2025, available at https://www.cfodive.com/news/deloitte-refunds-60k-report-ai-errors-australian-government-accounting/803321/

Amiram, Dan, Zahn Bozanic, James D. Cox, Quentin Dupont, Jonathan M. Karpoff, and Richard Sloan. 2018. Financial reporting fraud and other forms of misconduct: A multidisciplinary review of the literature. *Review of Accounting Studies* 23: 732-783.

Beatty, Anne, and Joseph Weber. 2006. Accounting discretion in fair value estimates: An examination of SFAS 142 goodwill impairments *Journal of Accounting Research* 44(2): 257-288.

Denis, David J. and Diane K. Denis. 1995. Performance changes following top management dismissals. *Journal of Finance* 50(4): 1029-1057.

Dong, Mengmind Michael, Theophanis T. Stratopoulos, and Victor Xiaoqi Wang. 2024. A scoping review of ChatGPT research in accounting and finance. *International Journal of Accounting Information Systems* 55: 1-29.

Haman, M., and M. Školník. 2024. Using ChatGPT to Conduct a Literature Review. *Accountability in Research* 31 (8): 1244–1246. https://doi.org/10.1080/08989621.2023.2185514.

Hanlon, Michelle, and Shane Heitzman. 2010. A review of tax research. *Journal of Accounting and Economics* 50: 127-178.

Hines, James R., and Eric M. Rice. 1994. Fiscal paradise: Foreign tax havens and American business. *Quarterly Journal of Economics* 109(1): 149-182.

Khanna, Tarun, and Krishna Palepu. 2000. Is group affiliation profitable in emerging markets? An analysis of diversified Indian business groups. *Journal of Finance* 50(2): 867-891.

Kraft, Arthur, Andrew J. Leone, and Charles Wasley. 2006. An analysis of the theories and explanations offered for the mispricing of accruals and accrual components. *Journal of Accounting Research* 44(2): 297-339.

Neumeister, Larry. 2023. Lawyers submitted bogus law created by ChatGPT. A judge fined them $5,000. *AP News* June 22, available at https://apnews.com/article/artificial-intelligence-chatgpt-fake-case-lawyers-d6ae9fa79d0542db9e1455397aef381c.

Skolnik, Sam. 2024. Lawyer sanctioned over AI-hallucinated case cites, quotations. *Bloomberglaw.com* November 26, available at https://news.bloomberglaw.com/litigation/lawyer-sanctioned-over-ai-hallucinated-case-cites-quotations.

Somasundaram, R. 2023. Step-by-step guide: How to use ChatGPT to write a literature review with prompts. *iLovePhD*, available at https://www.ilovephd.com/step-by-step-guide-how-to-use-chatgpt-to-write-a-literature-review-with-prompts/.

Stapleton, Andy. 2024. How to use ChatGPT to write a literature review. *Academia Insider*, available at https://academiainsider.com/use-chatgpt-to-write-a-literature-review/.

Wolf, Rachel. 2025. Federal judge chooses not to sanction lawyer who admitted using AI in mistake-filled brief. *Fox News* March 4, available at https://www.foxnews.com/us/federal-judge-chooses-not-sanction-lawyer-who-admitted-using-ai-mistake-filled-brief.

Vakilzadeh, Hamid, and David Wood. 2025. The development of a RAG-based artificial intelligence research assistant (AIRA). Forthcoming, *Journal of Information Systems*.

Wood, David A., and 327 coauthors. 2023. The ChatGPT artificial intelligence chatbot: How well does it answer accounting assessment questions? *Issues in Accounting Education* 38(4): 1-28.

Wood, David A. 2025. *Rewiring Your Mind for AI: How to Think, Work, and Thrive in the Age of Intelligence*. Technics Publications.

Wu, Qingyun, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyn Zhang, Shaokun Zhang, Liale Liu, Ahmed Awadallah, Ryen W. White, Doug Burgber, and Chi Wang. 2024. AutoGen: Enabling next-gen LLM applications via multi-agent conversations, *conference paper, Conference on Language Models (COLM).*

Table 1

Table 1 shows our results from 33 prompts of accounting research topics to ChapGPT 4o basic or free version ("4 Omni"), Panel A, ChatGPT 4o Plus (a subscription account offering an upgraded model), Panel B, Gemini 2.5 flash (the free AI model offered by Google), Panel C, and Perplexity, Panel D. We show the total number of papers cited broken out by correct citations ("real" papers), real papers with errors in the journnal name, real papers with other citation errors (incorrect year, volume, or issue), real but misleading or mischaracterized papers, and the number of paper citations where we cannot locate the actual paper ("fakes" or "hallucinations").

	Panel A		**Panel B**		**Panel C**		**Panel D**	
Generative AI Used	ChatGPT 4o base		ChatGPT 4o Plus		Gemini 2.5 flash		Perplexity	
Number of prompts	33		33		33		33	
Average papers cited per prompt:	10.1		6.7		12.3		8.2	
	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
Total papers cited	334		221		405		269	
Papers correctly cited	244	73.1%	99	44.8%	186	45.9%	190	70.6%
"Real" papers with errors:								
Incorrect journal name	13	3.9%	31	14.0%	48	11.9%	11	4.1%
Incorrect citation	17	5.1%	36	16.3%	79	19.5%	44	16.4%
Misleading	1	0.3%	13	5.9%	2	0.5%	15	5.6%
Total "real" but with errors	31	9.3%	80	36.2%	129	31.9%	70	26.0%
Total "real" papers	275	82.3%	179	81.0%	315	77.8%	260	96.7%
Nonexistent papers ("hallucinations")	59	17.7%	42	19.0%	90	22.2%	9	3.3%
	334	100.0%	221	100.0%	405	100.0%	269	100.0%

Table 2

Table 2 shows our results for a comparison of a no-option ChatGPT 4o prompt responses (Panel A), the "think longer" option (that is, "tool," Panel B), and the "deep research" tool (Panel C). We select four prompts: (1) "list prior accounting research papers on the causes and effect of CEO and CFO turnover on firm performance with citations and urls," (2) "List prior accounting research papers on taxes with citations and urls," (3) "List prior accounting research papers on fair value accounting with citations and urls," and (4) "List prior accounting research papers on tax avoidance with citations and urls."

	Panel A		**Panel B**		**Panel C**	
Tool Used	None		Think Longer		Deep Research	
Number of prompts	4		4		4	
Average papers cited per prompt:	9.8		9.0		11.0	
Average time needed per prompt	about 10 seconds		about 22 seconds		11 minutes 30 seconds	
	Number	Percentage	Number	Percentage	Number	Percentage
Total papers cited	39		36		44	
Papers correctly cited	26	66.7%	28	77.8%	36	81.8%
"Real" papers with errors:						
Incorrect journal name	5	12.8%	4	11.1%	1	2.3%
Incorrect citation	2	5.1%	2	5.6%	7	15.9%
Misleading	0	0.0%	0	0.0%	0	0.0%
Total "real" but with errors	7	17.9%	6	16.7%	8	18.2%
Total "real" papers	33	84.6%	34	94.4%	44	100.0%
Nonexistent papers (hallucinations)	6	15.4%	2	5.6%	0	0.0%
	39	100.0%	36	100.0%	44	100.0%

Table 3

Table 3 shows our results for a comparison of a no-option Gemini prompt responses (Panel A) and the "deep resaerch" option (that is, "tool," Panel B). Gemini does not offer a "think longer" tool. We select four prompts: (1) "list prior accounting research papers on the causes and effect of CEO and CFO turnover on firm performance with citations and urls," (2) "List prior accounting research papers on taxes with citations and urls," (3) "List prior accounting research papers on fair value accounting with citations and urls," and (4) "List prior accounting research papers on tax avoidance with citations and urls."

	Panel A		**Panel B**	
Tool Used	None		Deep Research	
Number of prompts	4		4	
Average papers cited per prompt:	11.5		26.3	
Average time needed per prompt	about 10 seconds		about 6 minutes	
	Number	Percentage	Number	Percentage
Total papers cited	46		105	
Papers correctly cited	23	50.0%	77	73.3%
"Real" papers with errors:				
Incorrect journal name	7	15.2%	1	1.0%
Incorrect citation	12	26.1%	19	18.1%
Misleading	0	0.0%	6	5.7%
Total "real" but with errors	19	41.3%	26	24.8%
Total "real" papers	42	91.3%	103	98.1%
Nonexistent papers ("fakes")	4	8.7%	2	1.9%
	46	100.0%	105	100.0%

Table 4

Table 4 shows our results for a comparison of a no-option Perplexity prompt responses (Panel A) and the "deep resaerch" option (that is, "tool," Panel B). Perplexity does not offer a "think longer" tool. We select four prompts: (1) "list prior accounting research papers on the causes and effect of CEO and CFO turnover on firm performance with citations and urls," (2) "List prior accounting research papers on taxes with citations and urls," (3) "List prior accounting research papers on fair value accounting with citations and urls," and (4) "List prior accounting research papers on tax avoidance with citations and urls."

	Panel A		**Panel B**	
Tool Used	None		Deep Research	
Number of prompts	4		4	
Average papers cited per prompt:	8.3		26.0	
Average time needed per prompt	about 10 seconds		1 minute 41 seconds	
	Number	Percentage	Number	Percentage
Total papers cited	33		104	
Papers correctly cited	27	81.8%	80	76.9%
"Real" papers with errors:				
Incorrect journal name	2	6.1%	6	5.8%
Incorrect citation	2	6.1%	8	7.7%
Misleading	0	0.0%	9	8.7%
Total "real" but with errors	4	12.1%	23	22.1%
Total "real" papers	31	93.9%	103	99.0%
Nonexistent papers ("fakes")	2	6.1%	1	1.0%
	33	100.0%	104	100.0%

Table 5

Table 5 shows our results for a comparison of Deep Research prompt responses for ChatGPT, Gemini, and Perplexity alongside prompt results for a newer AI research assistant, AIRA. We select four prompts: (1) "list prior accounting research papers on the causes and effect of CEO and CFO turnover on firm performance with citations and urls," (2) "List prior accounting research papers on taxes with citations and urls," (3) "List prior accounting research papers on fair value accounting with citations and urls," and (4) "List prior accounting research papers on tax avoidance with citations and urls."

	Panel A		**Panel B**		**Panel C**		**Panel D**	
AI Model	ChatGPT 4o		Gemini 2.5 Flash		Perpexity		AIRA	
Tool Used	Deep Research		Deep Research		Deep Research		n/a	
Number of prompts	4		4		4		4	
Average papers cited per prompt:	11.0		26.3		26.0		15.0	
Average time needed per prompt	11 minutes 30 seconds		about 6 minutes		1 minute 41 seconds		about 1 minute	
	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
Total papers cited	44		105		104		60	
Papers correctly cited	36	81.8%	77	73.3%	80	76.9%	52	86.7%
"Real" papers with errors:								
Incorrect journal name	1	2.3%	1	1.0%	6	5.8%	0	0.0%
Incorrect citation or duplicate	7	15.9%	19	18.1%	8	7.7%	2	3.3%
Misleading	0	0.0%	6	5.7%	9	8.7%	6	10.0%
Total "real" but with errors	8	18.2%	26	24.8%	23	22.1%	8	13.3%
Total "real" papers	44	100.0%	103	98.1%	103	99.0%	60	100.0%
Nonexistent papers ("fakes")	0	0.0%	2	1.9%	1	1.0%	0	0.0%
	44	100.0%	105	100.0%	104	100.0%	60	100.0%

Appendix: List of Query Prompts

1	Accounting research on CEO and CFO turnover.
2	Business research on CEO and CFO turnover.
3	List prior accounting research papers on the accruals anomaly.
4	List prior accounting research papers on earnings manipulation with citations.
5	List prior accounting research papers on accounting restatements with citations.
6	List prior accounting research papers on auditor changes with citations.
7	List prior accounting research papers on fraud with citations.
8	List prior accounting research papers on the F-score with citations.
9	List prior accounting research papers on accounting estimates with citations.
10	List prior accounting research papers on financial distress with citations.
11	List prior accounting research papers on auditor succession with citations.
12	List prior accounting research papers on small auditing firms with citations.
13	List prior accounting research papers on SEC enforcement releases with citations and urls.
14	List prior accounting research papers on financial statement forecasting with citations and urls.
15	List prior accounting research papers on the effect of the Sarbanes Oxley Act with citations and urls.
16	List prior accounting research papers on CFO turnover with citations and urls.
17	List prior accounting research papers on market efficiency with citations and urls.
18	List prior accounting research papers on artificial intelligence with citations and urls.
19	List prior accounting research papers on audit pricing with citations and urls.
20	List prior accounting research papers on social responsibility with citations and urls.
21	**List prior accounting research papers on taxes with citations and urls.**
22	**List prior accounting research papers on tax avoidance with citations and urls.**
23	List prior accounting research papers on corporate governance with citations and urls.
24	List prior accounting research papers on financial statement quality with citations and urls.
25	List prior accounting research papers on earnings quality with citations and urls.
26	List prior accounting research papers on measuring business risks with citations and urls.
27	List prior accounting research papers on investment with citations and urls.
28	List prior accounting research papers on conservatism with citations and urls.
29	**List prior accounting research papers on fair value accounting with citations and urls.**
30	List prior accounting research papers on intangible assets with citations and urls.
31	List prior accounting research papers on financial statement credibility with citations and urls.
32	List prior accounting research papers on the Dechow F-score with citations with citations and urls.
33	**List prior accounting research papers on the causes and effect of CEO and CFO turnover on firm performance with citations and urls**

Note: Bold prompts were prompts selcted for "think longer" (ChatGPT) and "deep research" (all AI models)