Literature search with Large Language Models: Insights for the social sciences

Jan Karem Höhne
DZHW, Leibniz University Hannover

Fernanda Alvarado-Leiton
University of Costa Rica

Abstract
Researchers are increasingly leveraging Large Language Models (LLMs) for writing and literature search, raising concerns about reference discrepancies (e.g., nonexistent articles). When LLMs generate fabricated references, these "hallucinations" destabilize epistemic trust and scientific integrity. We deployed a factorial study design across four LLMs, four prompt designs, and four social science research areas. Each prompt requested ten references and was applied to each LLM and research area with two prompting iterations, resulting in 1,192 references. We evaluated LLM-generated references to quantify hallucinations and shed light on their extent. Our results reveal that while hallucinations exist, their share is low (less than 7%). More concerning is that LLMs generate little reference variation (only 32% of unique references). A phenomenon that we name "Literature Diversity Deficit." Overall, literature search with LLMs can complicate knowledge curation, archiving, and transfer. The limited reference diversity may also amplify citation inequalities, homogenized literature, and reinforced epistemic bias.

Keywords: Academic writing, Citation accuracy, Fabricated references, Knowledge production, Literature Diversity Deficit, Scientific integrity

Introduction and theoretical background

Since the public launch of OpenAI's ChatGPT in late 2022, Large Language Models (LLMs) have elicited unprecedented public interest. Only two months after its launch, ChatGPT reached 100 million active users, a benchmark that took social media platforms, such as TikTok nine months and Instagram two years to reach (The Guardian, 2023). In July 2025, ChatGPT further expanded to reach about 10% of the global adult population, with practical guidance (e.g., tutoring and creative ideation), writing (e.g., production of emails and documents), and information seeking (e.g., searching for information about people and events) emerging as the three most common conversation topics that collectively account for about 80% of all conversations (Chatterji et al., 2025).

This document is a preprint and thus it may differ from the final version.

Following ChatGPT's launch, additional "general-purpose" LLMs, such as Google's Gemini, alongside open-weight LLMs, such as Meta's Llama, were rapidly made accessible to the public.

Parallel to this widespread public surge, the utilization of LLMs has unfolded rapidly within academia. Among researchers, LLMs offer substantial advantages by supporting them across technical tasks, such as generating analysis code (Nejjar et al., 2025), as well as cognitively demanding, theory-driven tasks, such as research ideas and hypotheses generation (Mullainathan & Rambachan, 2026; Zhou et al., 2024). As shown by Alvero et al. (2026), writing assistance appears to be one of the most frequently conducted tasks, with researchers frequently utilizing LLMs for grammar correction, language editing, and manuscript drafting. For example, Cao (2026) analyzed nearly 1.5 million research abstracts and estimated that LLM-based text modification increased from near zero in late 2022 to almost 25% by late 2025 (see also Liang et al., 2025). Importantly, the uptake of LLMs in the academic context is not restricted to researchers but also extends to university students. Analyzing self-reports from over 20,000 university students across 109 countries, Amoah et al. (2025) reported that more than 70% of the students utilize ChatGPT for their academic studies.

Beyond writing assistance, LLMs are increasingly utilized for literature search. However, this shift toward LLM-based literature search introduces some critical vulnerabilities, as LLMs have shown to generate hallucinated references (Boretti, 2026; Buchanan et al., 2024; Naddaf, 2026; Walters & Wilder, 2023). These hallucinated references appear plausible at first glance but include partially or entirely fabricated references with incorrect or nonexistent metainformation, such as incomplete author lists, fictitious article titles, nonexistent journals, or incorrect Digital Object Identifiers (DOIs) (Boretti, 2026).

Existing literature indicates that hallucinated references are widespread. For example, Walters and Wilder (2023) observed that between 18% and 55% of the LLM-generated references were completely fabricated. Among the real, non-fabricated references, between 24% and 43% of the references contained substantive metainformation errors, with numerical components, such as publication years, volumes, issues, and page numbers, being particularly prone to inaccuracies. Similarly, Buchanan et al. (2024) reported that when instructed to generate academic summaries across economic themes, about 30% of the references provided by ChatGPT were nonexistent. This share was higher for niche (or specific) than general research themes. Similarly, an audit of 2.5 million articles revealed that almost 3,000 articles included fabricated references and that the rates have climbed steeply since 2023 (Naddaf, 2026). This evidence serves as cautionary tales to researchers, as uncritical delegation of literature search to LLMs poses severe long-term risks to scientific integrity and knowledge curation, archiving, and transfer, potentially undermining trust in research and its processes.

Hallucinated references can be contextualized through the framework of "Quadruple Deception" proposed by Simon (2025). In this framework, the author describes four forms of deception: (1) deception regarding the ontological status of the interaction partner in the form of Artificial Intelligence (AI), (2) deception regarding the functional capacities of AI, (3) deception through the AI-generated content itself, and (4) deception arising from the opaque integration of AI into other software environments. From our perspective, hallucinated references can be directly linked to the

second and third forms of deception. For example, when researchers use LLMs for literature search, they may succumb to the second form of deception by misinterpreting probabilistic, pattern-predicting LLMs as knowledge- or information-retrieval tools. This misattribution of functional capacity enables the third form of deception, as LLMs can produce plausible yet hallucinated references with realistic-appearing titles, authors, and journals. These intersecting deceptions carry profound ramifications. For example, the proliferation of hallucinated references has the potential to pollute future research articles and reference databases, destabilizing epistemic trust and leaving researchers vulnerable to placing faith in deceptive outputs and eroding collective academic confidence.

To better understand the potential risks related to LLM-based literature search one can build on the concept of "Sealed Knowledges" by Lindemann (2025), articulating how LLMs alter access to information when used in the form of search engines. In contrast to "traditional" (scholarly) search engines, such as Google Scholar, which structure and rank search results into distinct links while preserving the original context, metainformation, and source for users' evaluation, LLMs tend to collapse this into a single synthesized output (or reference). Therefore, LLMs rather operate as a narrow, algorithmic filter or "seal" that compresses vast, complex, and context-dependent scholarly information into a highly fluent and seemingly authoritative reference output. This sealing mechanism aligns with Latour's (1987) concept of the "black box," wherein complex systems function such that only their inputs and outputs are visible, effectively obscuring the internal processes that produced them.

This structural compression creates two key epistemic risks relevant to LLM-based literature search. First, it potentially masks the situated origins of scientific knowledge through de-contextualized references that cannot reflect the underlying nature of complex scientific debates. Second, by design, chat-based LLM interfaces encourage users to prioritize convenience over critical thinking, making them less likely to verify details and sources or even question the generated output. This convenience encouragement relates to Mühlhoff's (2018) critique of digital disempowerment, particularly to his discussion of how user experience design strategies can shape user behavior and undermine critical engagement. In LLM-based literature search, among other things, this risks missing discoveries of unexpected yet valuable perspectives by narrowing the scope of literature reviews.

To address these concerns, an expanding body of literature has begun to empirically examine hallucinated references. Nonetheless, upon examination of the state of the art, we identified some important limitations that need to be addressed. First, existing studies often focus on single disciplines and rather specific research themes, such as biomedical science (Naddaf, 2026), computational linguistics (Sakai et al., 2026), economics (Buchanan et al., 2024), and maxillofacial surgery (Jain et al., 2025), offering limited insights into the broader social sciences. An exception is the study by Walters and Wilder (2023), who conducted a multidisciplinary analysis with a set of over 40 research themes spanning the natural sciences (e.g., differences between molten salt reactors and conventional nuclear fission reactors), humanities (e.g., quantitative methods in the analysis of English-language literary), and social sciences (e.g., impact of Brexit on the UK economy). These

studies consistently report hallucinated references. However, discipline-focused investigations within social sciences and their research themes remain rare.

Second, LLMs and their model versions emerge rapidly making prior evidence more challenging to interpret. Previous studies have shown that LLMs, such as ChatGPT and Gemini, produce hallucinated references at varying rates (see, for example, Buchanan et al., 2024; Cabezas-Clavijo & Sidorenko-Bautista, 2026; Walters & Wilders, 2023). However, because LLM architecture and training data undergo constant changes and updates (Chen et al., 2024), their performance and hallucination rates cannot be assumed to be static. Therefore, continuous auditing is key to evaluating whether the advancement of LLMs has meaningfully reduced the occurrence of hallucinated references or merely altered their manifestations.

Third, prompt designs have been shown to be associated with the extent of hallucinated references, although best design practices remain lacking. For example, Davis and Mahmoud (2026) showed how the type and level of guidance provided to the LLMs yields different rates of hallucinated references. By comparing four LLMs, the authors found that prompts providing detailed information on, for example, reference number, source type, and reference style were associated with higher rates of fabricated or inconsistent references, whereas less detailed prompts more frequently produced unsupported conceptual claims rather than hallucinated references (see also Niimi, 2025).

Fourth, existing studies also vary with respect to the type of task and reference generation constraints provided to LLMs. For example, some studies instructed LLMs to embed references within generated text (Walters & Wilder, 2023; Franzoni Velázquez et al., 2024) and other studies instructed LLMs to generate standalone reference lists (Franzoni Velázquez et al., 2024; Kim et al., 2025), while also differing in whether they restrict generation to specific bibliographic formats, such as journal articles (Kim et al., 2025).

Fifth, evidence suggests that LLM performance is better for general themes when compared to specific ones (Buchanan et al., 2024). An explanation for this phenomenon may simply lie in the amount of training data (e.g., the number of existing research articles) that is generally higher for general than specific research themes (see Davis & Mahmoud, 2026). However, so far, there is only a very small body of studies systematically investigating LLM-generated references across research areas that, for example, vary with respect to maturity and breadth.

To address these limitations, in this study, we conduct a comprehensive and systematic investigation of LLM-generated references within the social sciences. We focus on peer-reviewed journal articles as they are the primary benchmark of vetted academic scholarship (Reinhart & Schendzielorz, 2024). Specifically, we evaluate reference generation for the social sciences across four general-purpose LLMs, four prompt designs, and four research areas that systematically vary along the dimensions of research maturity and breadth. Put succinctly, we address the following three research questions (RQs):

To what extent do LLMs generate hallucinated references in social science literature search? (RQ1)

Can prompt design reduce hallucinated references in LLM-based social science literature search? (RQ2)

Are research areas differently susceptible to hallucinated references in LLM-based social science literature search? (RQ3)

Method

To investigate our three research questions, we considered four LLMs, four prompt designs, and four research areas, resulting in a 4 × 4 × 4 factorial study design. All research areas consisted of one research theme (e.g., social capital) for which we instructed the LLMs to provide 20 important "peer-reviewed journal articles." The generation process was split into two iterations of ten articles each. All references were manually inspected by a human coder to assess whether and to what extent they were hallucinated or not. In what follows, we initially provide information on the LLMs, prompt designs, and research areas (including research themes). We then outline the reference generation, coding procedure, and analytical strategy.

Large Language Models (LLMs)

In this study, we investigated four state-of-the-art LLMs: OpenAI's ChatGPT, Google's Gemini, Anthropic's Claude, and Mistral AI's Mistral.[2] While the first three models were developed and are provided by US-American companies, the last one belongs to a European-owned company. These models were selected as they represent "general-purpose" LLMs that are publicly accessible through chat interfaces, which facilitate the use of these tools for scientific writing and literature search. In addition, these models have been utilized in previous studies related to LLM-based literature search. Nonetheless, existing studies partially differ in terms of scopes (e.g., focusing on academic training), included disciplines (e.g., health science), and LLM tasks (e.g., writing of research papers) (Cabezas-Clavijo & Sidorenko-Bautista, 2026; Walters & Wilder, 2023).

Prompt designs

We developed four so-called "role-playing" prompts (Tseng et al., 2024) to investigate whether and to what extent different prompt designs result in different levels of reference hallucinations during literature search. The four prompt designs differ in the level and type of guidance provided to the LLMs. While the benchmark prompt (baseline condition) represents a general literature search request, the remaining prompts incorporate different strategies intended to improve generation results. First, the "admitting uncertainty" prompt asked the model to explicitly acknowledge when reference information cannot be verified. Second, the "constraint" prompt restricts recommendations to journal articles indexed in Scopus (https://scopus.com) and/or Web of Science (https://webofscience.com). Finally, the "reflective" prompt instructs the model to explain its selection and verification process. Importantly, we asked LLMs to exclusively provide references to peer-reviewed journal articles following the ASA (American Sociological Association) citation

[2] We set up accounts for each chat-based LLM on September 09, 2026, selecting a basic license plan for all LLMs under investigation to enhance comparability: ChatGPT (plan: GO for 8€ per month), Gemini (plan: Google AI Plus for 4,99€ per month), Claude (plan: Pro for 18€ per month), and Mistral (plan: Pro for 17,99€ per month). In addition, before reference generation we selected the highest performing model and prompt setting available and deactivated the memory feature (to mitigate the occurrence of carry-over effects).

style, requiring full reference disclosure (e.g., including publication year, volume, issue, page number(s), and DOI). Together, these prompt designs enable a holistic evaluation of whether and to what extent the type and level of LLM guidance is associated with the occurrence of hallucinated references. Table 1 lists all four prompt designs.

Table 1. Prompt designs and prompts for instructing LLMs

Prompt design	Prompt
Benchmark	You are a social scientist writing a research article on ["theme"]. Please provide 10 important peer-reviewed journal articles about ["theme"] following the citation style of the American Sociological Association (ASA).
Admitting uncertainty	You are a social scientist writing a research article on ["theme"]. Please provide 10 important peer-reviewed journal articles about ["theme"] following the citation style of the American Sociological Association (ASA). **Please explicitly state "I cannot verify this journal article" if you are not certain a journal article exists.**
Constraint	You are a social scientist writing a research article on ["theme"]. Please provide 10 important peer-reviewed journal articles about ["theme"] following the citation style of the American Sociological Association (ASA). **Please select only journal articles that are indexed in Scopus or Web of Science.**
Reflective	You are a social scientist writing a research article on ["theme"]. Please provide 10 important peer-reviewed journal articles about ["theme"] following the citation style of the American Sociological Association (ASA). **Please elaborate on the selection and verification of each journal article.**

Note. We explicitly stated the theme (e.g., "social capital") for which the LLMs were instructed to generate references. This is indicated by the brackets in each prompt. All prompts contain the same main instruction, and all instructions that go beyond the baseline prompt are highlighted in bold.

Research areas

To emulate the vast complexity and diversity of the social sciences, we selected four research areas that systematically vary along two dimensions: research maturity and research breadth. We defined research maturity as the temporal duration over which an area has been studied (established vs. contemporary), and research breadth as the conceptual scope of an area (general vs. specific). Together, these two aspects account for potential variations in literature search performance stemming from differences in training data volume (or literature abundance), terminology stability, and degree of disciplinary specialization (Davis & Mahmoud, 2026; Niimi, 2025).

The definition of the four research areas was based on two cut-off points: For the research maturity cut-off, we selected the year 2010 as the baseline threshold to differentiate between established and contemporary research areas. This year falls into a pivotal transition period with respect to the social sciences, reflecting the rise of a digital research landscape characterized by digitally mediated data, computational methods, and novel social phenomena (Lazer et al., 2009). A theme was classified as established if at least 50% of its total Google Scholar results were published

prior to 2010. Conversely, a theme was classified as contemporary if more than 50% of its results were published in 2010 or later. For the research breadth cut-off, we established a threshold of one million search results to distinguish between general and specific research areas. Specifically, a research theme yielding 1 million or more total results was classified as general, whereas a theme yielding fewer than 1 million results was classified as specific. Finally, we discussed research area definition and theme selection with expert colleagues at the sociological departments of our universities.

We crossed our two dimensions to obtain the following four research areas: Established-general (i.e., one million or more results with 50% before 2010), contemporary-general (i.e., one million or more results with 50% from 2010 and onwards), established-specific (fewer than one million results with 50% before 2010), and contemporary-specific (i.e., fewer than one million results with 50% from 2010 and onwards). Table 2 provides an overview of the research areas, selected themes, and descriptions of each theme.

To validate and classify our selected research themes in Table 2, we conducted an exploratory bibliometric analysis using Google Scholar (https://scholar.google.com) on July 23, 2026. For each research theme, we recorded both the total number of search results (by selecting the "any time" search option) as an indicator of research breadth, and the temporal distribution (by setting the time range to "2010 to 2026") as an indicator of research maturity. All search terms were enclosed in quotation marks and entered in lowercase (e.g., "social capital"), ensuring that Google Scholar would retrieve exact phrase matches rather than broad keyword combinations, independent of capitalization variants.

Reference generation

Reference generation was conducted in a narrow time frame, from September 14 to 25, 2026, to minimize possible noise introduced by potential model updates or changes during generation. The four LLMs under investigation were queried using an identical procedure. Within each LLM, the four prompt designs were administered in the same order: Benchmark, admitting uncertainty, constraint, and reflective. Each prompt requesting ten references was applied independently to each research theme with two prompting iterations, resulting in 20 LLM-generated references per research theme. This resulted in a total of 1,192 references.[3] The iterated prompt design aims to reflect upon the variability of the generative process of LLMs. To generate the references, we created new accounts for all LLMs and initiated a new chat session for each prompt design and iteration to mitigate potential carry-over effects (Appendix A provides an example screenshot of the reference generation process).

[3] Gemini generated 232 instead of the 320 references requested.

Table 2. Research areas, selected themes, and descriptions of each theme

Research areas	Themes	Theme descriptions
Established-general	Social capital	Social capital describes the resources and advantages embedded in social relationships, networks, and structures, and the concept has been studied across various aspects of social life, including educational attainment, cultural consumption, migration decisions, economic prosperity, and community well-being (Erices-Ocampo et al., 2025).
Contemporary-general	Big data	Big data refers to the collection, processing, and analysis of large-scale datasets characterized by, for example, high volume, velocity, and variety, which has transformed approaches to behavioral and social research (Lazer & Radford, 2017).
Established-specific	Satisficing	Satisficing is a concept related to rational choice that describes the tendency of individuals to reduce cognitive effort, particularly in decision-making contexts and survey participation (Krosnick, 1991).
Contemporary-specific	Digital well-being	Digital wellbeing concerns the relationship between digital technology use and individual psychological, social, and behavioral outcomes, including technology-related stress and healthy technology practices (Uslu, 2025).

Note. Social capital achieved a total of 3,450,000 Google Scholar search results with 1,890,000 before 2010. Big data achieved a total of 3,330,000 Google Scholar search results with 1,970,000 from 2010 and onwards. Satisficing achieved a total of 103,000 Google Scholar search results with 75,000 before 2010. Digital wellbeing achieved a total of 8,290 Google Scholar search results with 7,030 from 2010 and onwards. Google Scholar search was conducted on July 23, 2026.

In line with our prompts in Table 1, all LLMs were instructed to provide ten important peer-reviewed journal articles per iteration corresponding to the respective research themes. Only the initial response generated by the model was retained for analysis. If an LLM generated multiple alternative reference lists (e.g., "Option A" and "Option B") within a single response, only the first complete list ("Option A") was retained for analysis. Subsequent alternatives were also disregarded. No follow-up prompts or clarifications were requested, ensuring that all references were generated under identical conditions. All generated references were documented immediately after generation and archived in a coding database without modifications, except for title transformation into lower case letters, which was necessary for the subsequent coding procedure.

Coding procedure

All LLM-generated references were manually coded by a human coder. As LLMs generate references without a predefined "ground truth," the coding entails an initial identification step that assesses whether a reference exists or not. To verify each reference, the coder first searched for the Digital Object Identifier (DOI) in the official DOI-system (https://www.doi.org). If an LLM-generated reference did not contain a DOI or the respective DOI could not be identified (as it was incorrect or hallucinated), the coder conducted a query-based Google Scholar search as another attempt. In doing so, the title of the LLM-generated reference was copied-and-pasted into the search bar of Google Scholar and enclosed in quotation marks (e.g., "toward a unified conceptualization of social capital;" Erices-Ocampo et al., 2025). In this case, the coder searched Google Scholar and the verifiability of the reference was evaluated based on the title and metainformation, such as author and journal names. We considered information provided by official journal pages, JSTOR, and Google Books as reliable sources. We did not consider third-party .pdf files and "gray literature" as reliable sources. Based on this identification process, the coder categorized each LLM-generated reference as 1 "verifiable reference" or 0 "non-verifiable reference."

In the next step, the coder assessed the bibliographic type of verifiable references. More specifically, he evaluated whether an LLM-generated reference referred to a peer-reviewed journal article or other bibliographic materials (e.g., monographs, theses, book chapters, preprints, working papers, and blog posts). This was done through the DOI-system or Google Scholar. Based on the reference information, the coder then categorized each LLM-generated reference as 1 "peer-reviewed journal article" or 0 "non-peer-reviewed journal article."

Excluding all non-peer-reviewed journal articles from the subsequent coding, the coder then evaluated potential discrepancies with respect to the article's metainformation in the form of author name(s), titles, publication years, journals, volumes, issues, page numbers, and DOIs. Based on this evaluation, the coder annotated all metainformation using the following categories: 1 "correct" and 0 "incorrect" (including not provided necessary metainformation). Importantly, as not all peer-reviewed journal articles consist of a volume, issue, page number(s), and DOI, the coder was additionally instructed to assign 1 "correct" for those cases.[4]

[4] For example, if a reference missed or falsely included an author it was categorized as incorrect. However, if a reference missed an author's second name (e.g., "Krosnick, Jon" instead of "Krosnick, Jon A.") it was still categorized as correct.

For some prompt designs, further tailored coding was necessary. In particular, for the "admitting uncertainty" prompt design, the coder checked whether the LLMs admitted uncertainty regarding a reference by stating that they cannot verify a journal article (1 "admitted" or 0 "not admitted"). For the "constraint" design, the coder additionally checked whether a reference was indexed in Scopus and/or Web of Science (1 "indexed" or 0 "not indexed"). Finally, the coder examined the reasoning texts generated by the LLMs ("reflective" design) with the following categorization: 1 "valid reasoning" or 0 "invalid reasoning" (including no reasoning provided).[5] Figure 1 presents a flowchart of the coding procedure.

To evaluate the accuracy of the coding, a second independent coder manually coded a randomly selected subset of 10% of all unique references (n = 39). This resulted in an average agreement rate of 93.4%.

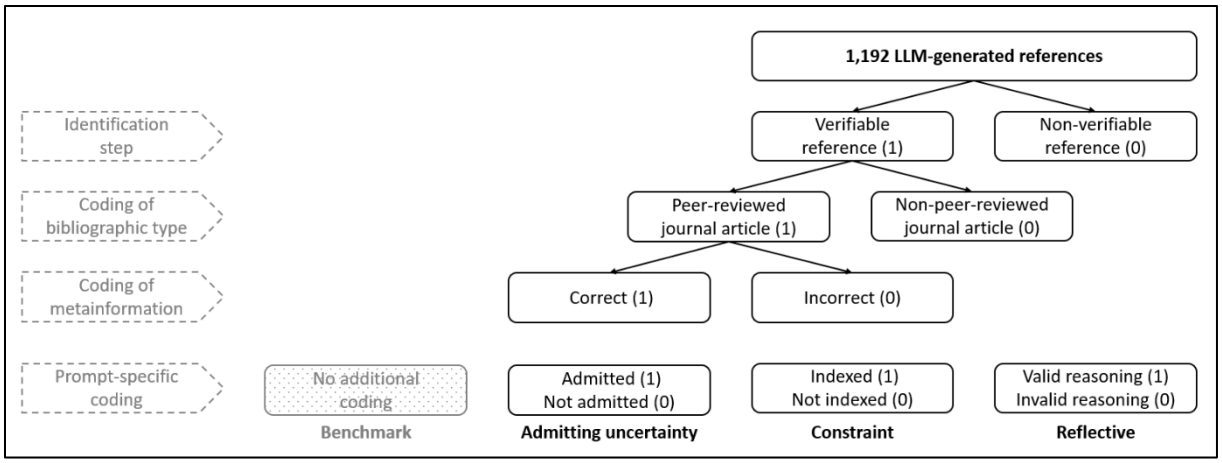


Figure 1. Coding flowchart

Analytical strategy

In a first step, we calculate descriptive statistics on the generated references across LLMs, prompt designs, and research areas. To this end, we report on the number of references generated and the number of unique references generated across all LLMs, including the overlap between the two prompting iterations. We also report percentages on the reference verifiability and whether a reference is a peer-reviewed journal article (called "pertinence"). Then, we report the correctness of the article's metainformation in the form of a "Metainformation Quality Index" (MQI) that is based on the author name(s), titles, publication years, journals, volumes, issues, page numbers, and DOIs. Subsequently, we report on the prompt-specific instructions related to the LLM-generated references: percentage of uncertainty admissions, percentage of references indexed in Scopus and/or Web of Science, and percentage of valid reasonings provided.

In total, we have three dependent variables for our subsequent analysis: our first dichotomous dependent variable is "verifiability," which is based on the reference identification process. Our second dichotomous dependent variable is "pertinence." Our third metric dependent variable is the MQI, which is an additive index counting the incidences of correctly generated

[5] Reasonings were categorized as valid when they were given at the reference level and referred to the aspects of selection and verification.

metainformation varying between 0 (no correct metainformation) and 8 (entirely correct metainformation). It is based on the following metainformation: author(s), publication year, title, journal, volume, issue, page number(s), and DOI.

Our analytical strategy aims to investigate our research questions through generalized linear models[6] for each of the three dependent variables. We use Maximum Likelihood Estimation (MLE). For our dichotomous dependent variables (verifiability and pertinence), we assume a binary family and logistic link function. For the MQI, in contrast, we use a tobit regression model (Tobin, 1958) to accommodate for the maximum value of the index and its asymmetric distribution.

With respect to our independent variables, we create dichotomous variables to indicate the type of LLM, prompt design, and research area. For LLMs, we create Gemini (1 = "Yes"), Claude (1 = "Yes"), and Mistral (1 = "Yes") with ChatGPT as reference. For prompt design, we create admitting uncertainty (1 = "Yes"), constraint (1 = "Yes"), and reflective (1 = "Yes") with benchmark as reference. Finally, for research areas, we create contemporary-general (1 = "Yes"), established-specific (1 = "Yes"), and contemporary-specific (1 = "Yes") with established-general as reference.

Results

Descriptive statistics

In total, the four LLMs generated 1,192 references. The number of references generated was consistent across LLMs (320 references per LLM), with only Gemini generating 232 out of the 320 references requested. Even though the LLMs generated such a large number of references, we found substantial redundancy. Specifically, the four LLMs generated only 385 (or 32.3%) unique references across prompt designs and research areas. On average, the two iterations yielded 14.1 unique references out of the 20 references possible. There was little variation between LLMs, prompt designs, and research areas (see Appendix B).

Next, we investigated the three main variables of interest. When it comes to verifiability, we were able to verify 1,114 (93.5%) of the references, which indicates a baseline hallucination rate of 6.5%. Regarding pertinence, we found that 1,029 (92.4%) of these verifiable references corresponded to peer-reviewed journal articles.[7] With respect to the quality of the generated metainformation, we obtained an overall average MQI score of 7.3 (out of 8 potential points) for the verifiable peer-reviewed journal articles. Appendix B reports more details on these descriptive statics.

However, it is important to note that Gemini refused to generate references following ASA style guidelines. Instead, it generated references following the style guidelines of the American Psychological Association (APA).

We found acceptable performance of LLMs regarding prompt compliance. For the prompt in which the LLM was asked to admit uncertainty, we encountered such declarations for 26 (10.5%) of the verified, peer-reviewed journal articles. Nonetheless, LLMs were more likely to admit uncertainty for non-verifiable references: 20 (33.9%) of these references were

[6] We accounted for the nested data structure using mixed-effects models. However, the inter-iteration variance was negligible and did not improve the model fit.

[7] When exclusively looking at the 385 unique references the picture slightly changes, as we obtained 340 (88.3%) of verifiable references and 305 (89.7%) of pertinent references.

flagged with an uncertainty declaration. When asked to constrain the references, LLMs generated 269 (97.8%) references that were indeed indexed in Scopus and/or Web of Science. Finally, when we asked the LLMs to reflect on the reason why they chose a reference we found that valid reasonings were provided for 208 (83.5%) references. Appendix C reports the prompt-specific coding outcome in more detail.

Reference verifiability

Table 3 summarizes the odds ratios from a logistic regression model with verifiability as dependent variable (LR $\chi^2(6) = 122.72$, p-value < 0.001). Note that references generated by ChatGPT were excluded from this model due to perfect separation, as ChatGPT achieved a 100% verifiability rate. In addition, the research area variables were collapsed into only two variables (1 = "Contemporary" with 0 = "Established" as reference), as only one reference was non-verifiable within the contemporary-specific condition (quasi-perfect separation).

Table 3. Parameters of a logistic regression model with verifiability as dependent variable

Independent variables	Odds ratio	Standard error	p-value
LLMs with Gemini as reference			
Claude	1.37	1.12	0.704
Mistral	0.04	0.03	< 0.001
Prompt design with benchmark as reference			
Admitting uncertainty	0.77	0.27	0.468
Constraint	1.34	0.51	0.449
Reflective	0.87	0.31	0.700
Research area with established as reference			
Contemporary	1.46	0.38	0.143
Observations	872		
Pseudo-R^2	0.23		

Note. LR $\chi^2(6) = 122.72$, p-value < 0.001. ChatGPT was excluded due to perfect separation. Research area variables were recoded due to instability in parameter estimation.

Turning to the results, the model parameters reveal performance disparities across LLMs (RQ1). With Gemini serving as the reference category, Mistral exhibits lower verifiability when it comes to its generated references. More specifically, the results indicate that the likelihood of a generated reference being verifiable is about 96% lower for Mistral than for Gemini. In contrast, there is no difference in reference verifiability between Claude and Gemini.

Variations in prompt design (RQ2) do not yield differences when it comes to the verifiability of the LLM-generated references. Relative to the baseline benchmark prompt, neither the admitting uncertainty prompt, the constraint prompt, nor the reflective prompt shows meaningful associations with the likelihood of producing verifiable references.

Finally, regarding research areas (RQ3), again, no difference is observed. References generated for contemporary research areas do not differ with respect to verifiability from those in established areas.

Reference pertinence

Table 4 summarizes the odds ratios from a logistic regression model with the pertinence of LLM-generated references (i.e., whether verified references correspond to peer-reviewed journal articles) as dependent variable (LR $\chi^2(6) = 66.13$, $p < 0.001$). For the same reasons described above, the specification of this model was done without ChatGPT and research areas were again collapsed into two variables.

Table 4. Parameters of a logistic regression model with pertinence as dependent variable

Independent variables	Odds ratio	Standard error	p-value
LLMs with Gemini as reference			
Claude	5.83	2.54	< 0.001
Mistral	0.49	0.13	0.007
Prompt design with benchmark as reference			
Admitting uncertainty	1.45	0.47	0.247
Constraint	2.74	0.98	0.005
Reflective	1.30	0.41	0.397
Research area with established as reference			
Contemporary	1.48	0.36	0.110
Observations	794		
Pseudo-R^2	0.12		

Note. LR $\chi^2(6) = 66.13$, $p < 0.001$. ChatGPT was excluded due to perfect separation. Research area variables were recoded due to instability in parameter estimation.

The odds ratios from the model reveal differences across LLMs when it comes to reference pertinence (RQ1). For example, the odds ratio for Claude is more than five times as high as that for Gemini, which serves as reference. In contrast, Mistral exhibits a lower likelihood of generating pertinent references. In fact, the odds ratio of generating pertinent references is about 51% lower for Mistral than for Gemini.

Regarding prompt design (RQ2), we only observe differences in reference pertinence for one of the prompt designs. Specifically, neither the admitting uncertainty prompt nor the reflective prompt shows an association with the likelihood of generating peer-reviewed journal articles. In contrast, prompting the LLM to constrain the reference generation to only peer-reviewed journal articles indexed in Scopus and/or Web of Science results in higher odds of obtaining such references.

With respect to our research areas (RQ3), references generated for contemporary research areas are not more likely to be peer-reviewed journal articles than those generated for established areas.

Metainformation quality

Finally, we examine the variables associated with the quality of reference metainformation among verified, peer-reviewed journal articles, measured through the MQI. To accommodate the upper limit of the index and its skewed distribution, we specified a Tobit regression model (Tobin, 1958): (LR $\chi^2(9) = 923.30$, $p < 0.001$). Unlike previous model specifications, all variables were included as described in the "Analytical strategy" section, as there were no separation issues. Table 5 summarizes the parameters of the Tobit regression model with MQI as dependent variable.

Table 5. Parameters of a tobit regression with MQI as dependent variable

Independent variables	Coefficients	Standard error	p-value
LLMs with ChatGPT as reference			
Gemini	-3.81	0.17	< 0.001
Claude	-1.15	0.16	< 0.001
Mistral	-2.87	0.17	< 0.001
Prompt design with benchmark as reference			
Admitting uncertainty	-0.32	0.12	0.006
Constraint	0.18	0.12	0.133
Reflective	0.21	0.12	0.095
Research area with established-general as reference			
Contemporary-general	0.12	0.12	0.336
Established-specific	0.44	0.12	< 0.001
Contemporary-specific	-0.10	0.11	0.380
Observations	1,029		
Pseudo-R^2	0.36		

Note. LR $\chi^2(9) = 923.30$, $p < 0.001$.

The model shows variation across LLMs in terms of metainformation quality (RQ1). All LLMs produce lower MQIs than ChatGPT, which serves as reference. This is indicated by the negative coefficients.

Regarding prompt design (RQ2), only the admitting uncertainty prompt diverged from the benchmark regarding MQI. On average, this prompt yielded slightly lower MQI values.

Finally, turning to research areas (RQ3), MQI is largely stable across conditions, with one notable exception. Relative to the established-general research area, which serves as reference, LLM-generated references within the established-specific research area yielded slightly higher values of the MQI.

Discussion and conclusions

The aim of this study was to investigate the capabilities of LLMs in the context of social science literature search. We did this by addressing three research questions: the extent to which LLMs generate hallucinated references (RQ1), whether prompt design can mitigate hallucinated references (RQ2), and whether research areas vary with respect to hallucinated references (RQ3). We employed a factorial study design across four LLMs, four prompt designs, and four research areas. Our results indicate that while hallucinated references persist, their prevalence remains low (less than 7%). In addition, we found that prompt design is not sufficient to prevent hallucinated references and that this phenomenon is uniform across research areas.

Regarding our first research question, our overall results reveal that hallucination rates are lower than reported in prior studies (Walters & Wilders, 2023). For example, ChatGPT flawlessly generated verifiable and pertinent references, whereas Mistral exhibited substantial drop-offs in verifiability, pertinence, and metainformation quality. This heterogeneity carries epistemic risks. When LLMs are treated as a homogenous or interchangeable technology, they potentially succumb to what Simon (2025) conceptualizes as the second form of deception, misinterpreting probabilistic language models as standardized, factual knowledge- and information-retrieval tools. Substituting one LLM for another may introduce reference and metainformation pollutions. As shown in this study, even among highly capable LLMs, errors in metainformation persist. Because these errors are embedded within otherwise authentic references, they easily evade superficial scrutiny, which can create a false sense of security while compromising the integrity of reference archiving, curation, and transfer across the social sciences.

With respect to our second research question, our results indicate that different prompt designs offer limited help for enhancing LLM-generated references. While the constraint prompt yielded a minor positive association with pertinence, it neither improved metainformation quality nor prevented the recirculation of references. This result suggests that prompt engineering is no panacea. One should be cautious not overestimating the value of additional LLM guidance, as it does not provide guardrails nor enforce scholarly rigor. Attempting to "engineer away" hallucinations through prompt designs does not appear viable, underscoring that human oversight remains indispensable rather than delegable to the models themselves. This highlights the necessity of "human-in-the-loop" approaches (Mellamphy, 2021), keeping scholars actively involved in the reference generation process. Indeed, this imperative is further amplified, as chat-based LLM interfaces are optimized for convenience, encouraging what Mühlhoff (2018) critiques as digital disempowerment. In the context of LLM-based literature search, digital disempowerment manifests itself when scholars uncritically accept references, forfeiting their scientific responsibility and thoroughness.

When it comes to our third research question, our results reveal that research areas, varying across temporal maturity and conceptual breadth, exhibit uniform susceptibility to hallucinated references and metainformation quality issues. Contrary to the assumption that specific or contemporary research areas with sparser training data would yield higher error rates (Buchanan et al., 2024), we observed general stability across research areas and themes. This unexpected uniformity suggests that LLM performance in literature search is not simply a function of literature abundance or data scarcity related to research areas. Whether searching within established-general or contemporary-specific areas, researchers seem to face a persistent risk of encountering low reference diversity.

Overall, our study highlights structural issues that should raise concerns about LLM-based social science literature search. By presenting a condensed subset of references as representative of a research area, LLMs function as a "black box" in the sense of Latour (1987), obscuring their complex internal processes and mechanisms. Additional to this black-boxing, problematic literature search results are deepened by deceptions related to perceptions of LLM capacities (Simon, 2025). For example, one may mistake LLMs for objective retrieval engines while wrongly assuming their outputs are neutral and intentional. In this way, pluralistic scholarly perspectives are compressed into what Lindemann (2025) terms "Sealed Knowledges," where the inherent complexity, partiality, and situatedness of scientific discourses are flattened and alternative voices are locked out. This knowledge sealing is particularly dangerous for students, as they are still in the process of acquiring scientific expertise, potentially lacking sufficient knowledge to identify concealed complexities or question LLM content that appears plausible at first glance (Bråten et al., 2026). To navigate these challenges, the social sciences must cultivate a critical understanding that resists interface-related convenience (see Mühlhoff, 2018).

Finally, and above all else, our results reveal a previously unaddressed vulnerability: the LLMs under investigation generate little reference variation. Out of the 1,192 LLM-generated references, only 385 (32%) references were unique. We name this phenomenon "Literature Diversity Deficit," indicating that LLMs are not an encompassing search tool, but operate as narrow algorithmic filters that recirculate a constrained reference subset. Reference concentration carries epistemological consequences, as this homogenization threatens to narrow the intellectual scope of the social sciences by potentially reinforcing echo chambers and sidelining marginal, critical, or emerging scholarship. In addition, this compression may introduce further inequalities into the scholarly community. Current research on LLM-generated references largely neglects the structural and socio-demographic characteristics of the resulting corpora. In consequence, it is of utmost importance that future research thus relates to the investigation of the authorial, institutional, and journal-level characteristics. As LLMs can be trained on biased data (Guo et al., 2026), the artificially narrow reference pool may be largely composed of literature originating from well-established institutions, Western-centric research teams, and specific socio-demographic groups. In addition, linking LLM-generated references with bibliometric databases, would also allow to evaluate the impact of journal-level characteristics, such as impact factors, open-access status, and ownership by the "Big 5" publishers (i.e., Elsevier, SAGE, Springer Nature, Taylor & Francis, and Wiley; Lendvai & Petra, 2026) on the reference generation processes.[8] Uncritical reliance on LLM-based literature search thus risks the concentration of academic capital, accelerating structural exclusion under the guise of technological neutrality.

References

Alvero, A. J., Stoltz, D. S., Stuhler, O., & Taylor, M. A. (2026). Generative AI in Sociological Research: State of the Discipline. Sociological Science, 13, 45-62. https://doi.org/10.15195/v13.a3

[8] Academic publishers have entered licensing agreements with AI companies, formalizing structured access to their peer-reviewed text databases for LLM training (Kwon, 2024).

Amoah, A., Asiama, R. K., & Kwablah, E. (2025). ChatGPT early usage among students: A global evidence of determinants. Development and Sustainability in Economics and Finance, 7, 100065. https://doi.org/10.1016/j.dsef.2025.100065.

Boretti, A. (2026). Hallucinations in generative AI: A threat to scholarly integrity and the urgent need for publisher-led academically supervised verification. Energy Research & Social Science, 136(104720). https://doi.org/10.1016/j.erss.2026.104720.

Bråten, I., Latini, N., & Strømsø, H. I. (2026). Exploring students' (mis)conceptions about ChatGPT-generated text: A qualitative study. Education and Information Technologies, 31, 4637–4658. https://doi.org/10.1007/s10639-026-13938-w

Buchanan, J., Hill, S., & Shapoval, O. (2024). ChatGPT hallucinates non-existent citations: Evidence from economics. The American Economist, 69(1), 80–87. https://doi.org/10.1177/0569434523121

Cabezas-Clavijo, Á., & Sidorenko-Bautista, P. (2026). Assessing the performance of 8 AI chatbots in bibliographic reference retrieval: Grok and DeepSeek outperform ChatGPT, but none are entirely accurate. Journal of Data and Information Science, 11(2), 342-354. https://doi.org/10.1515/jdis-2025-0326

Cao, L. (2026). Visualizing the adoption of Large Language Models across sociology subfields. Socius: Sociological Research for a Dynamic World, 12, 1–3. https://doi.org/10.1177/23780231261460657

Chatterji, A., Cunningham, T., Deming, D., Hitzig, Z., Ong, C., Shan, C., & Wadman, K. (2025). How People Use ChatGPT. https://cdn.openai.com/pdf/a253471f-8260-40c6-a2cc-aa93fe9f142e/economic-research-chatgpt-usage-paper.pdf

Chen, L., Zaharia, M., Zou, J. (2024). How is ChatGPT's behavior changing over time? arXiv. Harvard Data Science Review, 6(2), https://doi.org/10.48550/arXiv.2307.09009

Davis, K. & Mahmoud, Q. (2026). Citation constraints and reference hallucinations in Large Language Models. Proceedings of the The 39th Canadian Conference on Artificial Intelligence (Proceedings of Machine Learning Research), 318, 464-475. https://proceedings.mlr.press/v318/davis26a.html.

Erices-Ocampo, P., Lubbers, M. J., Adams, J. (2025). Toward a unified conceptualization of social capital. Annual Review of Sociology, 51, 171-190. https://doi.org/10.1146/annurev-soc-090924-032037

Franzoni Velázquez, A. L., Huerta, E., & Jensen, S. (2024). Retracting ChatGPT: completeness and relevance of academic references. Discover Education, 3(226). https://doi.org/10.1007/s44217-024-00333-1

Guo, Y., Guo, M., Su, J., Yang, Z., Zhu, M., Li, H., Qiu, M., & Liu, S. S. (2026). Bias in Large Language Models: Origin, evaluation, and mitigation. Electronics, 15(9), 1824. https://doi.org/10.3390/electronics15091824

Jain, A., Nimonkar, P., & Jadhav, P. (2025). Citation integrity in the age of AI: evaluating the risks of reference hallucination in maxillofacial literature. Journal of Cranio-Maxillofacial Surgery, 53(10), 1871-1872. https://doi.org/10.1016/j.jcms.2025.08.004.

Kim, E., Kipchumba, F., & Min, S. (2025). Geographic Variation in LLM DOI Fabrication: Cross-Country Analysis of Citation Accuracy Across Four Large Language Models. Publications, 13(4), 49. https://doi.org/10.3390/publications13040049

Krosnick, J. A. (1991). Response Strategies for Coping with the Demands of Attitude Measures in Surveys. Applied Cognitive Psychology, 5, 213–236.

Kwon, D. (2024). Publishers are selling papers to train AIs — and making millions of dollars. Nature News. https://www.nature.com/articles/d41586-024-04018-5#author-0

Latour, B. (1987). Science in action: How to follow scientists and engineers through society. Cambridge, MA: Harvard University Press.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L-., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., & Van Alstyne, M. (2009). Computational social science. Science, 323, 721–723. https://doi.org/10.1126/science.1167742

Lazer, D., & Radford, J. (2017). Data ex machina: Introduction to Big Data. Annual Review of Sociology, 43, 19-39. https://doi.org/10.1146/annurev-soc-060116-053457

Lendvai, G. F., & Petra, A. (2026). Artificial intelligence in academic practices and policy discourses across 'Big 5' publishers. Research Evaluation, 35, rvag004. https://doi.org/10.1093/reseval/rvag004

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., & Zou., J. (2025). Quantifying large language model usage in scientific papers. Nature Human Behaviour, 9, 2599–2609. https://doi.org/10.1038/s41562-025-02273-8

Lindemann, N. F. (2025). Chatbots, search engines, and the sealing of knowledges. AI & Society, 40, 5063-5076. https://doi.org/10.1007/s00146-024-01944-w

Mellamphy, N. B. (2021). Humans "in the Loop?" Human-centrism, posthumanism, and AI. Nature and Culture, 16(1), 11–27. https://doi.org/10.3167/ nc.2020.160102.

Mullainathan, S., & Rambachan, A. (2026). Science in the age of algorithms. In A. K. Agrawal, E. Brynjolfsson, & A. Korinek (Eds.), The Economics of Transformative AI (pp. 337–366). Chicago: University of Chicago Press.

Mühlhoff, R. (2018). Digitale Entmündigung und User Experience Design: Wie digitale Geräte uns nudgen, tracken und zur Unwissenheit erziehen. Leviathan, 46(4), 551–574. https://doi.org/10.5771/0340-0425-2018-4-551

Naddaf, M. (2026). Surge in fake citations uncovered by audit of 2.5 million biomedical-science papers. Nature, 653(8115), 658-659. https://doi.org/10.1038/d41586-026-00748-w

Nejjar, M., Zacharias, L, Stiehle, F., & Weber, I. (2025). LLMs for science: Usage for code generation and data analysis. Journal of Software: Evolution and Process, 37(1), Article e2723. https://doi.org/10.1002/smr.2723

Niimi, J. (2025). Hallucinations in bibliographic recommendation: Citation frequency as a proxy for training data redundancy. arXiv. https://doi.org/10.48550/arXiv.2510.25378

Reinhart, M., & Schendzielorz, C. (2024). Peer-review procedures as practice, decision, and governance—the road to theories of peer review. Science and Public Policy, 51(3), 543–552. https://doi.org/10.1093/scipol/scad089

Sakai, Y., Kamigaito, H., & Watanabe, T. (2026). HalluCitation Matters: Revealing the impact of hallucinated references with 300 hallucinated papers in ACL conferences. arXiv. https://doi.org/10.48550/arXiv.2601.18724

Simon, J. (2026). Generative AI, quadruple deception and trust. Social Epistemology, 40(1), 101-115, https://doi.org/10.1080/02691728.2025.2491087

The Guardian (2023). ChatGPT reaches 100 million users two months after launch. https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

Tobin, J. (1958). Estimation of relationships for limited dependent variables. Econometrica, 26 (1), 24–36. https://doi.org/10.2307/1907382

Tseng, Y. M., Huang, Y. C., Hsiao, T. Y., Chen, W. L., Huang, C. W., Meng, Y., & Chen, Y. N. (2024). Two tales of persona in LLMs: A survey of role-playing and personalization. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 16612-16631). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.969

Uslu, O. (2025). Understanding digital wellbeing: impacts, strategies, and the path to healthier technology practices. Discover Social Science and Health, 5, 145. https://doi.org/10.1007/s44155-025-00259-5

Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. Scientific reports, 13(1), 14045. https://doi.org/10.1038/s41598-023-41032-5

Zhou, Y., Liu, H., Srivastava, T., Mei, H., & Tan, C. (2024). Hypothesis generation with large language models. In Proceedings of the 1st Workshop on NLP for Science (NLP4Science) (pp. 117-139), Miami (USA). https://doi.org/10.18653/v1/2024.nlp4science-1.10

Appendix A

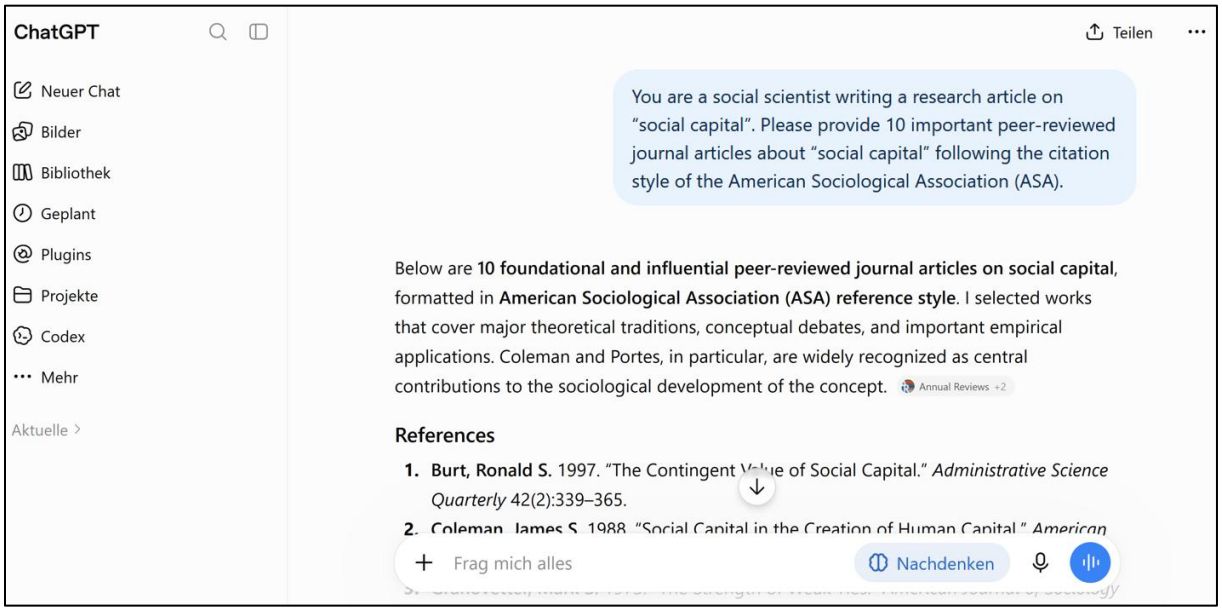


Figure A1. Illustration of reference generation through the interface of OpenAI's ChatGPT
Note. We employed the benchmark prompt (German interface).

Appendix B

Table B1. Descriptive statics across LLMs, prompt designs, and research areas

	ChatGPT	Gemini	Claude	Mistral
Number of generated references	320	232	320	320
Number of unique references	120	167	85	139
Percentage of verifiable references	100.0	98.7	99.1	77.5
Percentage of pertinent (peer-reviewed) references	100.0	88.2	97.8	79.4
Mean of MQI	8.0	6.1	7.8	6.9
	Benchmark	Admitting uncertainty	Constraint	Reflective
Number of generated references	306	291	303	292
Number of unique references	175	172	169	164
Percentage of verifiable references	93.8	92.1	95.1	92.8
Percentage of pertinent (peer-reviewed) references	89.6	92.5	95.5	91.9
Mean of MQI	7.3	7.2	7.4	7.4
	Established-general	Contemporary-general	Established-specific	Contemporary-specific
Number of generated references	314	292	299	287
Number of unique references	99	111	95	80
Percentage of verifiable references	92.7	89.0	92.6	99.7
Percentage of pertinent (peer-reviewed) references	90.0	90.8	92.4	96.2
Mean of MQI	7.2	7.4	7.5	7.3

Note. MQI stands for Metainformation Quality Index.

Appendix C

Table C1. Prompt-specific coding results in percentages

	Admitting uncertainty	Constraint	Reflective
LLM	Admitted	Indexed	Valid reasoning
ChatGPT	0.0	100.0	100.0
Gemini	20.5	89.7	46.7
Claude	0.0	100.0	77.9
Mistral	37.0	100.0	100.0