UDC 004.8:[37:001.891]

Kateryna Osadcha
Doctor of Pedagogical Sciences, Professor, post-doctoral researcher
Vrije Universiteit Brussel, Brussels, Belgium
leading researcher, Institute for Digitalisation of Education of the NAES of Ukraine, Kyiv, Ukraine
ORCID ID 0000-0003-0653-6423
Kateryna.Osadcha@vub.be, k.osadcha@iitlt.gov.ua

Maryna Osadcha
MA in Education, Lecturer
Bogdan Khmelnitsky Melitopol State Pedagogical University,
Zaporizhzhia, Ukraine
Mito Kiryo High School, Mito, Ibaraki, Japan
ORCID ID 0009-0007-6750-1853
niketa007@gmail.com

Natalia Shumeiko, PhD in Pedagogical Sciences, lecturer
Bratislava University of Economics and Business, Bratislava, Slovak Republic
ORCID ID 0000-0002-7859-7519
natalia.shumeiko@euba.sk

Volodymyr Proshkin
Doctor of Pedagogical Sciences, Professor, lecturer in maths
Rotherham College, Rotherham, UK
ORCID ID 0000-0002-9785-0612
Volodymyr.Proshkin@rnngroup.ac.uk

AI CHATBOTS IN RESEARCH PRACTICE: POSSIBILITIES FOR SEARCHING AND ANALYSING SCIENTIFIC SOURCES

Abstract. In modern research practice, AI chatbots are gaining recognition as effective digital assistants that facilitate the search and analysis of scientific literature. In this context, the article presents a comprehensive analysis of the features of using AI chatbots in research practice, particularly for searching and analysing scientific sources. The research methodology is based on a mixed approach that combines a desk review of scientific sources with a comparative assessment of results from 5 AI chatbots (ChatGPT, Microsoft Copilot, Gemini, Claude, Perplexity AI) across 8 criteria. The analysis used quantitative and qualitative metrics to assess AI chatbot responses to the same set of research queries by experts from four fields of knowledge (Engineering, Natural Sciences, Humanities, Social Sciences). A comparative analysis of AI chatbots showed their different capabilities and limitations in working with scientific sources. ChatGPT is the most balanced tool, with high accuracy and objective responses, but it does not cover many sources and is prone to interpretation. MS Copilot is distinguished by the speed of response generation and the sufficient relevance of its sources but is inferior in the objectivity of its analysis. Gemini provides sufficient correspondence with the topic and objectivity in response, but the responses are often excessively long and less accurate in analysis. Claude demonstrates the highest indicators of relevance and objectivity by working with a small number of sources, yet makes almost no generalisations. Perplexity AI is the leader in terms of the volume and quality of the source base, forms generalisations well, but is characterised by the lowest relevance of sources and objectivity of response. The results showed that all the studied AI chatbots demonstrate a high level of formal reliability of sources, but differ significantly in the relevance of sources, analytical depth, and the ability to interpret the scientific content of articles adequately. AI chatbots are effective tools for initial understanding of a research topic, rapid collection of sources for analysis and preliminary structuring of the material. However, they cannot fully carry out a deep and critical analysis of scientific sources on a specific research question.

Keywords: Artificial Intelligence; AI chatbots; Research practice; Evaluation of AI tools; AI-powered literary source search; AI analysis of academic sources.

© Kateryna Osadcha, Viacheslav Osadchyi, Natalia Shumeiko, Volodymyr Proshkin, 2026.

1. INTRODUCTION

The problem statement. In today's reality, we observe the rapid spread of generative AI (GenAI) across various areas of human activity [1], particularly its integration in education [2] and science [3]. Scientific research is an activity that requires careful, precise, and scrupulous work. This work requires considerable time. In particular, this work is time-consuming when scholars search for and analyse literature and recent publications in scientific databases. Importantly, before the rapid spread of GenAI and the release of ChatGPT 3.5 in 2022, researchers used academic databases (e.g., Scopus, Web of Science, PubMed, ERIC), search engines (e.g., Google Scholar), and multifunctional portals (e.g., ResearchGate) to locate literature. These, one might say, traditional literature search methods (e.g., bibliographic databases and search engines) have several limitations: complex query formulation, time-consuming analysis of large and often irrelevant result sets, duplicate detection and removal, and the need for extensive critical analysis of sources for qualitative research, despite the availability of numerous supporting bibliometric tools. With the increase in the number of scientific publications, both due to the digitisation of printed works and the growing publication activity, literature reviews take longer. PhD students sometimes spend a year analysing literary sources. This view is supported by researchers [4], who report that the average time from formulating the main idea of research to its publication is 67 weeks, approximately 1 year, and often 1 year and 3 months. Another investigation, which involved a literature review, concludes that only the librarian's help (i.e., an interview process, strategy development, search strategy translation, documentation, deliverables, writing methodology, and instruction) takes an average of 26.9 hours of working time. It should be noted that this is only a technical part, which does not include the months a PhD student spent reading literature and scientific sources, screening thousands of titles (i.e., books and dictionaries), finding data for the dissertation, and writing the dissertation text. All this proves that PhD students and experienced scientists face the same issue: the need to navigate the vast array of information to find the necessary literature without spending much time on it.

Effective management of global scientific information is currently one of the main challenges for the entire academic community, because the dynamics of scientific discoveries in any field depend directly on the speed of literature processing.

Analysis of recent studies and publications. We organised the literature review into three key dimensions to provide a comprehensive overview of the research issues.

Transformation of the SLR methodology: from the 'executor' to the 'architect'

The introduction of GenAI has significantly transformed the way systematic literature reviews (SLRs) are conducted. Traditionally, SLRs involved the manual processing of large datasets, but the process has since evolved toward a semi-automated approach. Researchers referred to this shift as the era of 'AI-enabled literature reviews' (AILR). In this new paradigm, the role of the researcher shifted from merely being an 'information processor' to that of the 'architect' of search strategies [5]. Empirical evidence supported the high efficiency of such collaboration, highlighting significant time savings and methodological optimisation [6]. A study reported that specialised GenAI-based systems achieved 89–97% accuracy in the screening and data extraction stages. These two stages are interrelated: first, the identified sources are filtered to select relevant ones (screening), and then specific facts or indicators are extracted from these sources to create a structured knowledge base (extraction). This effectiveness made AI performance comparable to that of human experts [7]. These findings aligned with the authors' approach in categorising modern tools into two groups: generative chatbots (such as ChatGPT and Claude) and specialised platforms (like Elicit, Consensus, and Scite.ai) [8].

Investigators proposed a structured three-level model for AI integration in research comprising preparation, initiation, and analysis. This model systematised the process of source selection [9]. Additionally, automation facilitated the development of 'living systematic reviews' that could be updated in real time. For instance, an AI agent scans databases for relevant keywords. When a new article is published, the AI analyses it to determine if it meets the research criteria. If it does, the article is added to the researcher's database. If the new article alters the study's overall understanding, the system alerts the researcher, and the 'living review' is updated accordingly [10]. Researchers [11] highlighted that the effectiveness of these tools heavily relied on the researcher's skills in 'prompt engineering'. This refers to the ability to interact with AI effectively, treating it as an assistant rather than merely a random text generator.

The performance paradox: balancing speed and scientific accuracy

While the integration of large language models (LLMs) offered significant advantages, it also entailed risks that have caused a 'performance paradox'. The essence of this paradox is that increased processing speed often leads to a reduction in the depth of analysis. One of the key challenges is the issue of 'hallucinations', where the model generates inaccurate or fabricated information. An experiment conducted by the authors [10] demonstrated that ChatGPT struggled to produce accurate bibliographic lists, often citing non-existent articles.

A four-level rating system (Grades A–F) was developed to evaluate the quality of bibliographic references. The results of their experiment were significant: in evidence-based medicine tasks, ChatGPT achieved only 0.5% accuracy in identifying relevant articles compared with an expert-curated reference list. This means the tool missed nearly 99.5% of the necessary sources and, in many cases, generated a substantial number of non-existent publications (rated Grade F as 'hallucinations'). Consequently, ChatGPT was deemed unsuitable for conducting systematic reviews that do not rely on external search databases. In contrast, Bing AI demonstrated a success rate of around 40% [12]. In addition to factual inaccuracies, there was also the issue of superficiality. Researchers conducted a comparative study and found that even advanced models like Claude AI, when provided with full texts, excelled primarily at summarising but struggled with deep synthesis and generating original insights [13]. Nonetheless, this technology held significant potential for inclusivity. GenAI tools could effectively assist non-native English speakers in overcoming language barriers and enhancing their academic writing style [14]. However, it is essential to engage in critical reflection to prevent over-reliance on these technologies [15].

Ethical boundaries and the crisis of authorship

The widespread availability of AI tools brought to light heated ethical debates. Chatbots cannot be used fully by the authors of scientific works because they are incapable of taking responsibility for the accuracy, integrity, and originality of the work, which is a necessary condition for authorship [16]. Researchers characterised AI models as 'stochastic parrots' that reproduce statistical patterns without understanding their content, since they only compute the conditional probability of the next word, ignoring the concept of truth. Such a mechanism creates the threat of plagiarism, a hidden form of plagiarism in which the intellectual contributions of the authors of the training data are dissolved within algorithms and presented in the generated text without proper references to the sources [17]. A survey confirmed these concerns: educators were concerned about the loss of the 'natural character' of scientific discourse and the possible degradation of researchers' critical thinking [18]. In response to the ethical challenges posed by the uncontrolled use of generative models, the scientific community has proposed new regulatory mechanisms. In particular, authors argued for the need to institutionalise the 'AI scrutiny' stage (expert evaluation of AI-generated content) in the review procedure. This process involves not only checking for plagiarism but also a targeted evaluation of the text for specific markers of generation: logical breaks, 'hallucinations' (fake links), and

unnatural language patterns, which allows the separation of original scientific work from synthetic work [14]. In parallel, in pedagogical and research contexts, researchers justified the transition to the use of 'reflection logs' (AI Reflection Logs). These are documented reports in which the authors detail each step of the interaction with the algorithm, from the formulation of prompts to the critical evaluation and editing of the results obtained. This approach deconstructs the concept of cheating (academic fraud, understood as the hidden delegation of cognitive efforts to the machine), transforming it into a transparent practice. Instead of blind copying, the researcher demonstrates intellectual engagement in data verification, thereby contributing to the development of academic integrity and digital literacy [19].

Summarising the analysed sources, it can be argued that the future of scientific research lies in the plane of hybrid intelligence. This is the concept of synergistic interaction, in which AI assumes the function of high-speed processing of large datasets (screening, extraction, primary systematisation). At the same time, the human researcher performs the indispensable functions of strategic planning, value judgment, and ethical control. In such a system, AI does not act as a substitute for the author, but as a powerful cognitive amplifier; its work is entirely subordinate to the scientist's critical thinking.

The research goals. Considering the above, this paper aims to investigate how AI chatbots can assist in performing research, particularly the search and analysis of scientific literature, at the initial stage of scientific exploration.

To achieve this aim, we set three research questions (RQ):

RQ1: What scientific publications analysing the capabilities of AI chatbots in terms of their potential in searching and analysing relevant scientific papers on research topics have already been published?

RQ2: What distinguishing features do ChatGPT, Microsoft Copilot, Gemini, Claude, and Perplexity AI have for the detailed search and analysis of the scientific literature?

RQ3: Can AI chatbots be a reliable tool for searching and analysing scientific literature?

Thus, this study presents a comparative analysis of 5 AI chatbots (ChatGPT, Microsoft Copilot, Gemini, Claude, Perplexity AI), focusing on their capabilities in searching for scientific sources. The research contributes to understanding the capabilities and limitations of these AI chatbots. In addition, this study continues the discussion of the ethical and methodological aspects of AI use in scientific practice.

2. THEORETICAL BACKGROUND

The theoretical background of the research of the functionality of AI chatbots in a literature search and conducting a literature analysis is rooted in a landmark theory and diverse approaches, particularly: (1) an 'information retrieval theory' [20]; (2) a systematic literature review [21]; (3) critical thinking concepts [22]; (4) an analysis of scientific works on the role of AI in scientific research [23]; (5) ways to assess AI chatbots [24], and (6) a sources about the AI chatbots under study [25].

The information retrieval theory, as an essential concept applied in this research, served as a guideline for determining criteria for evaluating how effectively AI chatbots can find relevant sources for scientific research. This study compares the performance of AI chatbots with traditional search engines, aiming to understand the relevance of AI chatbots' responses in terms of accuracy, reliability, and other criteria.

A systematic review methodology is applied to describe a standardised approach to literature review. This approach serves as a benchmark for assessing whether AI chatbots can support and automate various analysis stages of the literature review, structure, and analyse generalising sources. Moreover, this approach generally serves as an adequate tool to conduct reviews of scientific literature in compliance with academic standards.

The critical thinking concepts, as a cognitive basis for assessing the reliability, logic, and validity of information in research, help ensure the verification of the results provided by AI chatbots, identify errors or incorrect references, distinguish facts from assumptions, and distinguish relevant sources from a broader range of sources. Furthermore, these concepts allow for more accurate human (researchers) control and interpretation of the results.

Publications about the role of AI in scientific activity describe how AI transforms scientific practice, how AI modifies methods for working with scientific knowledge and data, how AI affects scientists' productivity and how AI places new emphasis on issues of academic integrity. These scientific publications serve as a basis for discussing AI chatbots' use not only as a technical tool but also as a socio-ethical phenomenon.

Publications on methods of evaluating AI chatbots' responses are valuable for us as they provide us with methodological guidelines for analysing AI chatbots' responses and a basis for selecting criteria for evaluating AI chatbots' responses to questions about literature search and analysis. Sources exploring AI chatbots' features, capabilities, functions, and characteristics (particularly for use in scientific practice) are useful for our scientific work.

The concepts of 'Academic Integrity and Responsible AI Use' are essential for the theoretical understanding of AI use in scientific practice. These concepts are interrelated and underpin the ethical application of AI, ensuring transparency, reliability, and trust for research results. Basic principles of 'Academic Integrity in the AI era', such as honesty, transparency, avoiding plagiarism, correct citation of sources, and responsibility for research results, are declared in 'Recommendation on the Ethics of Artificial Intelligence' [26], 'Guidelines for ethical use and acknowledgement of large languages in academic writing' [27], and other documents. The responsible AI use concept aligns with the principles that are outlined in 'Ethics guidelines for trustworthy AI' [28]. These principles are as follows: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; and accountability. Therefore, academic integrity provides the fundamental values of science and responsible AI use. These principles are crucial for the current AI era. In symbiosis, they provide a theoretical framework that underscores the importance of transparency, responsibility, and critical thinking.

Our choice of AI chatbots (ChatGPT, MS Copilot, Gemini, Claude, Perplexity AI) is motivated by the intention to focus on their distinct AI-powered capabilities [29]. Different approaches to AI use are applied in the functioning of these chatbots: ChatGPT – universal; Copilot and Gemini – integrated (in Microsoft and Google); Claude – explanatory; Perplexity AI – search. Let's consider their general characteristics and features for application in scientific practice in more detail.

2.1. ChatGPT

It is important to note that '*the development of ChatGPT was driven by the desire to create a highly sophisticated and versatile AI language model capable of assisting in various tasks, including text generation, translation, and data analysis'* [30]. ChatGPT is a deep learning-based chatbot, a text generation model based on the Transformer architecture, that can learn and understand human language, engage in dialogue, and interact in context. '*It can learn and understand human language, engage in dialogue, and interact in context. It can also complete tasks such as writing papers, stories, translating languages, and writing code. It can provide people with efficient and convenient intelligent services'* [31]. Scientists characterise ChatGPT as a multifunctional tool applicable to various stages of scientific research. There are different AI-powered models. For instance, the Research Librarian model proved to be a weak tool for finding scientific articles, as it generates incorrect link descriptions and even invents links.

GPT-4 can provide helpful feedback aligned with the generally accepted standards of modern research practice. As a Data Generator, ChatGPT can be useful for generating data because it can realistically model real data. However, GPT's ability to create functional hypotheses may depend on its ability to replicate existing results. As a Novel Data Predictor, ChatGPT cannot predict new data [32]. ChatGPT's deep research mode currently provides deep source search, data synthesis, and comprehensive report generation capabilities. The more complex Agent Mode helps perform multi-step tasks from start to finish, allowing the researcher to plan, act, and adapt to context, which is essential for complex scientific projects. Customising personalisation in ChatGPT provides the ability to store context, for example, according to the researcher's field of activity, customise style (scientific, formal, analytical), and use memory to track previous queries and topics.

2.2. MS Copilot

Microsoft Copilot Chat (MS Copilot) is a chatbot based on GenAI that uses the Microsoft Prometheus model, which integrates modern large language models (especially OpenAI). It is now not only a single product but also a modular ecosystem that operates on different language models (LLMs) using Microsoft software (Windows operating system, MS Office, MS Office 365, Edge browser) [33]. MS Copilot follows the principles of responsible AI. It can search the Internet, provide concise summaries of sources, offer useful advice and feedback, and provide accurate answers or help with creative tasks. It can maintain a meaningful conversation and help improve language skills [34]. MS Copilot is highly effective for creating quizzes on a specific topic, generating unit tests in programming, optimising code, and creating content for slides on a specific topic. At the same time, its ability to detect a logical error in a given article and to create a mental map to illustrate concepts is limited [35]. For scientific purposes, MS Copilot has several modes: 'Think Deeper' – advanced reasoning for complex questions, 'Deep Research' – deep search with detailed reports and links, 'Study' (GPT-5) – learning with hints and steps for deep reasoning or quick response depending on the task, 'Smart Mode' (GPT-5) – fast and deep thinking depending on the functions. With the 'Memory & Personalisation property', this AI chatbot can remember the user's scientific interests, communication style, priorities, and more. This allows users to get more relevant answers in long-term projects.

2.3. Gemini

Gemini models are built on top of Transformer decoders, with architecture and model optimisation that enable stable training at scale and optimised inference on Google's Tensor Processing Units. They are trained to support 32k context length, employing efficient attention mechanisms (e.g., multi-query attention). The new multimodal models, Gemini (Ultra, Pro, and Nano), exhibit remarkable capabilities across image, audio, video, and text understanding. They are suitable for applications ranging from complex reasoning tasks to on-device, memory-constrained use cases [36]. The study's [37] findings highlight Google Bard's (Gemini) systematic utility, demonstrating its role in efficient document organisation, retrieval, and the creation of research articles. According to the author, the tool's versatility lies in its application across diverse studies, thereby advancing cognitive skills and aligning with contemporary educational frameworks. Gemini 3 currently has a Deep Research tool that researchers can use for multi-stage and in-depth collection, analysis and synthesis of information on complex topics. Guided Learning can also be used for scientific purposes, helping users develop a deep understanding of the topic and critical thinking skills. The Deep Think model helps solve complex scientific tasks thanks to multi-step reasoning.

2.4. Claude

Claude is a family of state-of-the-art large language models developed by Anthropic. Claude 4.5 introduces three models [38]: 1) Claude Sonnet – productive for coding tasks, able to perform tasks related to the entire life cycle of development from the search for architectural solutions to testing and vulnerability detection; 2) Claude Haiku – positioned as the fastest and smartest Haiku model with almost advanced performance, has the function context awareness, enabling the model to track its remaining context window throughout a conversion; 3) Claude Opus – combines maximum capabilities with practical performance, is skilled in reasoning, coding and complex problem solving tasks, has low levels of unwanted behaviour [39]. Claude has 5 response writing styles: Normal – universal, quick answers; Learning – explanations through examples, questions, and hints; Concise – short, clear answers; Explanatory – detailed explanations with argumentation; Formal – academic and official answers. It is also possible to customise the user style, which will allow you to adapt it to specific user tasks. In the article [13], the authors concluded that Claude surpasses humans in the task of creating a literature review by rapidly generalising points of agreement across sources, but has limitations in synthesis, providing incomplete citations and false information. Claude can identify key topics and sources and demonstrates the potential to expand and accelerate research workflows when used responsibly alongside human scientists.

2.5. Perplexity AI

Perplexity AI is an AI-powered answering engine for fast, trustworthy research. It combines a live web search with multiple leading AI models to deliver up-to-date answers, backed by verifiable citations. Perplexity helps explore new topics, summarise content, or delve into complex questions [40]. Perplexity AI uses several AI models (the Sonar family) and several external LLMs (GPT-5.x, Claude4.x, Gemini 3 Pro, Grok 4.x) [41], among which the system automatically distributes requests depending on the request, selected mode, and subscription. The scientific community views Perplexity AI primarily as a hybrid tool that combines the capabilities of LLMs with those of real-time search engines. This distinguishes it from the classical AI-powered chatbots (such as ChatGPT), because it adheres to principles of source transparency and academic integrity. In particular, it searches educational collections and Wiley resources, allowing users to find academic resources quickly and supporting in-depth research with the ability to cite relevant sources. Perplexity AI has an academic filter (an academic mode) that prioritises results from scholarly databases, journals, and reputable academic publications, filtering out non-academic or general web sources. This ensures that the responses are grounded in research and scholarly consensus [42]. The research mode in Perplexity AI allows the researchers to conduct in-depth research based on their own built research plan and source analysis, with results that can be exported as .pdf, .docx, or markdown and converted to a Perplexity page. In addition, Perplexity AI can personalise the researchers' profile, so the researchers can specify their role as a teacher/researcher and describe their scientific interests, which allows them to receive responses in an academic style relevant to their interests.

An overview of the features of the selected AI-powered chatbots regarding their usefulness for scientific purposes is presented in Table 1.

Table 1

Features of AI-powered chatbots valuable for scientific practice

AI-Chatbots	Strengths	Limitations	Typical use cases	Features which are useful for research
ChatGPT	Versatility, cross-platform availability	Limitations of the free version, dependence on subscriptions	Content systematization, text writing, creative tasks, programming	Deep Research Mode, Agent Mode, and personalisation settings
MS Copilot	Integration with Windows, Office, and Edge	Dependence on the Microsoft Ecosystem	Assistance, document handling, office tasks	Think Deeper Mode, Deep Research Mode, Study Mode, Smart Mode (GPT-5), Memory & Personalisation
Gemini	Integration with Google search and office services Google	Dependence on the Google ecosystem	Analytics, data processing, information search	Deep Research and Guided Learning tools, Deep Think model
Claude	Focus on safety, ethics, and long contexts	Not very popular, has limited access	Research, analytical work	Formal Writing Style, Claude Opus model
Perplexity AI	Strong search functionality, source verification, emphasis on reliable references	Less versatile, weaker creative capabilities	Factchecking, research, fast internet search	Academic Focus, Research Mode, and personalisation

3. METHODOLOGY

A mixed methodology was chosen for the study, including experimental research design, determination of criteria for evaluating AI chatbots, quantitative and qualitative analyses conducted by experts based on the selected criteria.

3.1. Experimental research design

A single set of requests for 5 chatbots across various disciplines (Engineering, Natural Sciences, Humanities, and Social Sciences) was applied in the study:

1. Request for Engineering disciplines (Query #1): 'Review current research on the application of machine learning to optimise power consumption in computer networks.'
2. Inquiry for Natural Sciences (Query #2): 'Prepare a review of the literature of the last 5 years on the use of software for modelling climate change and forecasting extreme weather events.'
3. Request for the Humanities (Query #3): 'Review current research on the application of artificial intelligence to develop educational materials for university courses in the Humanities.'
4. Request for Social Sciences (Query #4): 'Prepare a literature review on the role of AI chats in shaping public opinion and disseminating information on social networks.'

3.2. Determination of criteria for evaluating the responses of AI chatbots

Building on the analysis of scientific publications that proposed criteria for evaluating AI chatbot responses [43] – [45], we opted for 8 criteria (Table 2) for evaluating AI chatbot

responses in literature searches and analyses. Our primary objective was to determine how accurate, relevant, and capable these responses were, and how they present the content in a structured form.

Table 2

Criteria for evaluating the responses of AI chatbots

No.	Criteria	Characteristics	Indicator (units of measurement)
1.	Relevancy	Source relevance indicator	Percentage of sources in the AI chat responses that are relevant to the stated research issue (%)
2.	Correspondence	Correspondence of the response to the topic of the request	Complete (2), Partially (1), Incomplete (0)
3.	Accuracy	Existence of sources	Percentage of sources that exist (%)
4.	Completeness	Adequate coverage of sources	Quantity of sources covered
5.	Depth	Availability of generalisations, conclusions, identification of trends, etc.	Yes (1) / No (0)
6.	Structuredness	Logical organisation of the content / material	Logical (clear topics, subtopics, chronology) (1) / Illogical (without clear structure) (0)
7.	Objectivity	Consistency between the brief sources analysis by AI chatbots and the text of the article	Percentage of sources in the AI chatbot response whose analysis by the AI chatbot matches the text of the cited article (%)
8.	Ethicality	Warning about possible errors and/or limitations in the generated response	Provides a warning (1) / Does not provide a warning (0)

Four experts from diverse fields (i.e., Engineering, Natural Sciences, Humanities, and Social Sciences) were engaged in assessing AI chatbot responses. Each expert assessed the responses according to the criteria. The assessment results are presented in the tables below (Tables 3–8). Evaluating AI chatbots' responses and their practical usefulness enabled researchers to assess the quality of AI-generated output. The evaluation includes criteria that clarify the potential of AI chatbots for analysing literature.

4. RESULTS

To get responses to queries for different subjects, we used free (without payment, 'clean') accounts for each AI chatbot under research (new accounts without a history of previous queries and without a description of the user's profile), and Deep Research mode or its analogue for each AI chatbot. Each request was made once without clarification, except when the AI chatbot stopped generating. In ChatGPT, the model that was used was OpenAI o3. Correspondingly, in MS Copilot – GPT-5, in Gemini – Gemini 3, in Perplexity AI – a specific combination of models that is automatically selected by the system for optimal performance of complex tasks [46]. To get responses from Claude, we used the Formal style of the Sonnet 4.5 model because the Claude Opus mode is only available in Claude Pro. For transparency and reproducibility, all original chatbot responses used in this study have been collected and made publicly available on Zenodo [47].

Next, we present the responses of AI chatbots and their analysis. The responses of the first chatbot (ChatGPT) are presented in Table 3.

Table 3

The assessment of ChatGPT's responses to the Queries, in accordance with the criteria

No	Criteria/ Query No.	1	2	3	4	An average
1.	Relevancy	66.7%	75%	11.1%	38.5%	47.8%
2.	Correspondence	2	1	0	1	1

3.	Accuracy	100%	100%	100%	100%	100%
4.	Completeness	12	8	9	13	11
5.	Depth	1	1	1	1	1
6.	Structuredness	1	1	1	1	1
7.	Objectivity	75%	63%	77.8%	77%	73%
8.	Ethicality	0	0	0	0	0
Additional parameters						
9.	Number of words	1444	1320	2357	2161	1802

Chat GPT does not add any warning information about errors/mistakes or restrictions to each response, but at the bottom of the website (chatgpt.com), there is a warning common to all requests. Also, when sharing a response via link, it provides a warning that the conversation may contain personal information. ChatGPT provides information on the query processing time, the number of queries and the number of sources. For some responses, ChatGPT adds illustrations. At the end of the response, ChatGPT provides a general summary and a brief analysis of the sources. While answering, the reasoning process is displayed step by step. ChatGPT allows exporting a response to a .docx format file. The report provides information on the number of sources processed but cites references for a much smaller number. ChatGPT, even in deep research mode, may cite non-scientific sources (e.g., media, blogs, online publications).

In general, it is possible to assert the average quality of the search for scientific sources, 100% accuracy across all 4 queries. The indicators of structuredness (=1) and depth (=1) are also very high. However, the correspondence rate indicator (=1) is average. The indicator of information completeness is not very high. We should also note the excessive 'cost' of collecting information: an average of 96 search queries, 6 minutes, 15 seconds, and ≈1800 words per response. It can be defined as 'undue research, which is too much,' and 're-writing' without significant completeness and relevance. The low average relevance rate (47.8%) indicates noticeable problems with focus on the topic and alignment with the query's theme. The ethicality indicator is also of concern. However, this does not imply that the responses are unethical. Rather, the responses lack explicit clarification regarding which elements may be subject to limitations or potential inaccuracies. This challenges scientific work, requiring additional verification from the researcher.

The assessment of Copilot's responses to the Queries is given below (Table 4).

Table 4

The assessment of Copilot's responses to the Queries, in accordance with the criteria

No	**Criteria/ Query No.**	**1**	**2**	**3**	**4**	**An average**
1.	Relevancy	61.5%	63%	35%	41.4%	50.2%
2.	Correspondence	2	2	1	2	1.8
3.	Accuracy	100%	100%	100%	100%	100%
4.	Completeness	26	46	28	29	32
5.	Depth	1	1	1	1	1
6.	Structuredness	1	1	1	1	1
7.	Objectivity	65.4%	71.7%	75%	65.5%	69%
8.	Ethicality	0	0	0	0	0
Additional parameters						
9.	Number of words	1261	1761	1532	1692	1562

Copilot does not add any warnings about errors/mistakes or restrictions to each response, but also at the bottom of the website (copilot.microsoft.com), there is a warning common to all requests. Copilot provides information about the request's processing time. The list of sources is provided separately in a format that is not convenient for copying. When converting a request into a .docx document, the sources are presented as notes. To create a request, Copilot does not ask additional questions; it generates a response immediately. While responding, Copilot displays, step by step, how the reasoning process unfolds. Copilot allows users to export responses in Word format. And finally, Copilot offers the steps for working with the response. For example,

'If you'd like, I can compile a literature matrix (methods, datasets, metrics, gains, limits) or draft a slide deck tailored to your environment (data-centre vs. RAN vs. IoT). What part of the network are you most interested in - cloud/DCN, 5G/6G RAN/transport, SDN/NFV, or edge/WSN?'

The peculiarity of Copilot lies in the process of creating a literary analysis. It is essential to note that the responses combine text (the essence of the response) with information from 2-3 sources, which makes it more difficult to realistically assess the objectivity of the response. For scientific purposes, it is important to follow the essence of the cited article more closely during literary analysis. Excessive creativity can distort the essence. It should also be noted that Copilot uses not only Internet resources but also GitHub resources, datasets, and Wikipedia articles to generate responses. Compared to ChatGPT, Copilot links not only to articles from arxiv.org and mdpi.com, but also to ieeexplore.ieee.org, researchgate.net, and springer.com.

Overall, the evaluation of Copilot's responses leads to the following conclusions. Half of the sources (relevancy / an average 50.2%) that Copilot offered corresponded to the request topic. It can be concluded that Copilot can find relevant sources but does not always accurately filter information, and in complex or interdisciplinary queries, relevance can fall below 50% (35% in a query about the Humanities). Copilot works well and aligns with topics when there are clear keywords, but it can 'expand' context, which reduces the accuracy of source selection. Copilot responses were fully or partially relevant to the query topic, suggesting a correct understanding of the query and the ability to maintain focus, but overgeneralisation led to departure into adjacent aspects. All the sources Copilot cited exist, and no sources were fabricated. Copilot provides more sources than ChatGPT but striving to cover a topic broadly does not always mean high relevance or objectivity. In all cases, Copilot demonstrates generalisation and a desire to identify trends, indicating the ability to draw analytical conclusions. Copilot's overall style is to give clear, structured, and readable answers. Copilot generally interprets sources correctly, but can sometimes overgeneralise, make not entirely appropriate assumptions, and combine sources into a single thesis. In deep research mode, Copilot did not warn about potential errors or limitations in its answers, but in normal mode, it often adds such warnings and notes about inaccurate interpretations when there is a lack of information. Perhaps, Copilot did not do this because of the academic style, or because of a lack of resources. Copilot provides slightly longer answers than ChatGPT, which aligns with its approach to handling academic queries. The processing time for requests varies depending on their complexity and the number of sources processed.

Table 5 below presents an analysis of Gemini responses according to the criteria.

Table 5

The assessment of Gemini's responses to the Queries, in accordance with the criteria

No	Criteria/ Query No.	1	2	3	4	An average
1.	Relevancy	40%	55.6%	30%	54%	44.9%
2.	Correspondence	2	1	1	1	1.25
3.	Accuracy	100%	100%	100%	100%	100%

4.	Completeness	30	54	40	50	44
5.	Depth	1	1	1	1	1
6.	Structuredness	1	1	1	1	1
7.	Objectivity	80%	73.7%	72.7%	67.4%	73.5%
8.	Ethicality	0	0	0	0	0
Additional parameters						
9.	Number of words	3028	2996	2632	3113	2942

Gemini, like previously analysed AI chatbots, does not add a warning about errors or restrictions to each response. However, at the bottom of the website gemini.google.com, there is a statement that Gemini can make mistakes. This chatbot first offers a plan for how it can consider the topic. If necessary, the user can review and edit the research plan in detail. During the response, the process of reasoning is shown step by step. Gemini does not provide information on processing time, the number of sources, or the number of queries. Unfortunately, Gemini involves for the analysis not only scientific articles in the deep research mode but also online sources (internet articles, blogs, media content), which reduces the scientific value of the results. Moreover, Gemini tends to over-generalise or form sentences that are far from the meaning of the context of the article that Gemini is analysing. The response time can exceed 10 minutes. Chatbot functions allow the user to export a response to a Google Docs format file.

Gemini demonstrates high-quality responses according to the criteria of accuracy (100%), objectivity (73.5%), depth, and structuredness. However, it has significant issues with relevance to the topic and ethical warnings (as we have identified them). According to experts, only about half of the sources (44.9%) actually relate to the query topic. It indicates instability in the selection of sources. Besides, sources on related topics are sometimes used. The correspondence indicator (1.25) yields a sufficient but not ideal result for scientific research, indicating a risk of deviation from the topic. It can be critical when conducting research. Accuracy at the level of 100% means the absence of hallucinations in the form of fictional articles or authors, which is essential for scientific works. In terms of source coverage, the completeness is average: the chatbot responses are sufficient, but the review is not exhaustive. Regarding the depth criterion, Gemini provides generalisations, conclusions, and analytical elements that demonstrate its ability to work not only at the level of facts but also at the level of interpretation. The responses of this chatbot are logically organised, with structure, subdivisions, and a precise, logical sequence of presentation. Most short analyses of sources align with the article content. Still, the reinterpretation of the Gemini text is sometimes so profound that it is difficult to find some correspondence with the source. The lack of ethical warnings in the response itself makes it less safe for academic use. Hence, Gemini can be used for analytical reviews, structured texts, generalisations, and preliminary source analysis. However, it is necessary to check the relevance of sources, monitor their significance to the topic, and most importantly, avoid relying on Gemini's answers as the only source of truth.

Table 6 below presents an analysis of Claude's responses according to the criteria.

Table 6

The assessment of Claude's responses to the Queries, in accordance with the criteria

No	**Criteria/ Query No.**	**1**	**2**	**3**	**4**	**An average**
1.	Relevancy	65%	53.3%	50%	-*	56.1%
2.	Correspondence	1	1	1	1	1
3.	Accuracy	100%	100%	100%	100%	100%
4.	Completeness	20 (13)**	30 (5)**	30(19)*	0*	20
5.	Depth	0	0	0	0	0

6.	Structuredness	1	1	1	1	1
7.	Objectivity	92.3%	80%	88%	-*	86.8%
8.	Ethicality	0	0	0	0	0
Additional parameters						
9.	Number of words	1396	2 996	1530	1306	1807

**Sources are not provided*

***Claude provides more sources at the beginning of the response (for example, 20 sources), but there is a different number of sources in the text itself (for example, 13 sources are provided).*

Claude's biggest downside is that it allows for only 2 queries every 4 hours. In order to send more queries (messages), the user should upgrade Claude to the Pro version (or Max). Claude, in each response, does not add a warning about the presence of errors or restrictions, but at the bottom of the website claude.ai, there is a warning common to all requests. In the free version of Claude, the user can only copy text. It is better to do so in the browser, not through the tool itself, because this way it is possible to save links to sources. In the response, Claude neither concludes nor provides a summary, nor does it include additional information, such as query processing time or the number of searches. In some responses, Claude provided a short response, and in some responses, Claude added a detailed review of the literature without reference to sources in a separate window.

The relevance indicator (56.1%) indicates that a substantial proportion of sources in the responses only partially align with the query's declared theme. Therefore, the search and selection of sources for analysis needs clarification. The correspondence between the response and the query theme is also partial. The accuracy (100%) indicates that all given sources are real, that is, there are no fictional references. Source completeness is low, especially given that many sources are not included in the response text itself as references to the article's specific text. As for depth, the indicator is 0, indicating that Claude does not make generalisations or conclusions. The response rather states the facts and interprets the source, but it does not always succeed. After all, in some cases of interpretation of the source text, it is difficult to determine to what extent Claude adheres to the essence. Claude's responses sometimes take a structured form, especially in the 4th response sample (a different text from previous responses is presented in the form of a structured Literature Review). The high objectivity rate (86.8%) is due to the limited number of sources cited in the answer. References are incorporated into the text as short summaries that generally match the article's content. The text without references shows Claude's verbosity but not the accuracy and objectivity of the answer, which are important for scientific research. Additionally, Claude does not warn about limitations or potential errors in each response, which is a significant drawback.

Claude does not use fictional sources and structures its responses well. The generated text corresponds closely to the cited article. However, its responses have an average indicator of relevance to source selection and correspondence to the topic, and a low indicator of depth of analysis and completeness of sources. His ethical rules regarding response templates are imperfect.

Table 7 below presents an analysis of Perplexity AI responses according to the criteria.

Table 7

The assessment of Perplexity AI responses to the Queries, in accordance with the criteria

No	Criteria/ Query No.	1	2	3	4	An average
1.	Relevancy	41.4%	43%	27.3%	33.1%	36.2%
2.	Correspondence	1	2	1.5	1.5	1.5

3.	Accuracy	100%	100%	100%	100%	100%
4.	Completeness	87 (41*)	173 (62*)	125 (30*)	132(29*)	129
5.	Depth	1	1	1	1	1
6.	Structuredness	1	1	1	1	1
7.	Objectivity	56.1%	61.3%	60%	69%	61.6%
8.	Ethicality	0	0	0	0	0
Additional parameters						
9.	Number of words	2541	2317	1464	3497	2455

**Perplexity AI provides more sources at the end of the response (for example, 87 sources), but there is a smaller number of sources in the text itself (for example, 41 sources are provided).*

Perplexity AI's most significant downside is that users with free access can submit only 3 in-depth queries per day. Perplexity AI, unlike other AI chatbots, neither provides warnings about errors or restrictions nor provides information about the query processing time or the number of attempts, like ChatGPT. However, Perplexity AI reports the number of processed sources and provides the sequence of actions (steps) it went through to generate the response. In the response, Perplexity AI provides an Executive Summary and Conclusion. Also, a note on the research methodology is provided. Sometimes Perplexity AI provides a short list of sources in the response and a separate list of used references. It is essential to note that generalisations and conclusions made by Perplexity AI are informative and qualitative. In research mode, Perplexity AI does not offer additional questions to continue or clarify the response at the end of the response, unlike in the standard mode. Like other AI chatbots, Perplexity AI can combine 2 or more sources in its analysis, making it harder to identify the correspondence between the source content and the analysis. It is noteworthy that Perplexity AI can use sources not only in English but also in Chinese, Japanese, and even Ukrainian.

The instability of the Perplexity AI deep research mode is evidenced by the average relevance score (36.2%) and the fluctuation of relevance scores from 27.3% to 43% in other responses. This is consistent with the developers' observation that a specific combination of models can be selected for each query. Perplexity AI selects sources accurately, but many of these sources do not correspond to the researcher's query. An average correspondence indicator of 1.5 is relatively high. That means that the responses were exactly on the topic and only partially went beyond the scope of the query, revealing tangential aspects. All the sources that Perplexity AI cited actually exist (accuracy =100%). Perplexity AI draws from many sources (an average of 129), especially compared to other AI chatbots. This confirms its ability to collect from a broad and diverse base of sources. The biggest plus of Perplexity AI is that it will generate the Executive Summary and the Research Conclusion. Sometimes, Perplexity AI even generates the Research Methodology, reflecting the analytical nature of the response rather than just a list of facts. Perplexity AI presents information logically, divides it into structured blocks, headings, and subheadings, and aligns with the chronology or applies a thematic approach. It is one of its most substantial advantages. The average indicator objectivity (61.6%) is not very high, and in some cases, this AI chatbot partially simplifies, overgeneralises, or exaggerates the article content. Perplexity AI does not provide warnings about potential errors, inaccuracies, or limitations, which may be a weakness from the perspective of academic standards and responsible use. The fact that Perplexity AI provides a large amount of text (an average of 2455 words) can be an advantage for researchers but a disadvantage for those seeking an accurate, short response.

In the summary table below (Table 8), we summarised all characteristics of all AI-powered chatbots that have been analysed in this study.

Table 8

Summary Table: The assessment of AI-powered chatbots' responses to the Queries, in accordance with the criteria

No	Criteria/ an AI chatbot	Chat GPT	Copilot	Gemini	Claude	Perplexity AI
1.	Relevancy	47.8%	50.2%	44.9%	56.1%	36.2%
2.	Correspondence	1	1.8	1.25	1	1.5
3.	Accuracy	100%	100%	100%	100%	100%
4.	Completeness	11	32	44	20	129
5.	Depth	1	1	1	0	1
6.	Structuredness	1	1	1	1	1
7.	Objectivity	73%	69%	73.5%	86.8%	61.6%
8.	Ethicality	0	0	0	0	0
Additional parameters						
9.	Number of words	1444	1562	2942	1807	2455

So, we chose 5 AI chatbots for the assessment (ChatGPT, Copilot, Gemini, Claude, Perplexity AI) and used 8 basic criteria and an additional parameter provided by the AI chatbots (number of words). Summarising the analysis conducted in this study, we can provide an assessment of the relevance, correspondence, accuracy, completeness, depth, structure, objectivity, and ethicality of the responses. The data in the summary table (Table 8) provide a general estimate of AI-powered chatbots' responses for each criterion.

Relevancy. Claude has the highest relevancy (56.1%), which is explained by a small number of proposed sources. Therefore, Claude selects these sources with high accuracy. Unlike Claude, Perplexity AI is, despite a large number of sources, much less relevant (relevancy = 36.2%).

Correspondence. Copilot (correspondence = 1.8) and Perplexity AI (correspondence = 1.5) correspond best to the topic of the query. At the same time, ChatGPT and Claude are less aligned with the query's theme.

Accuracy. All AI chatbots considered and assessed in this study showed 100% accuracy in deep learning mode for sourcing. That means there were no fictional sources, unlike in older versions of AI chatbots.

Completeness. Perplexity AI provides a large quantity of sources (an average of 129), which is helpful for researchers who want to explore tangential topics, but not for finding more specific responses. ChatGPT (completeness = 11) and Copilot (completeness = 44) work with fewer sources, but their responses are often more structured.

Depth and structuredness. The indicator of depth for all assessed AI chatbots is '1', except for Claude, which is '0', confirming AI chatbots' capability to draw conclusions, make generalisations, and highlight tendencies. Perplexity AI is especially indicative in this aspect. All analysed AI chatbots in this study, to some extent, structure their responses into logical blocks, chapters, and subchapters.

Objectivity. The most challenging thing for the experts of this study was to determine an indicator of objectivity – the correspondence of the analysis made by the chatbot with the real content of the source. Most often, the answers to the chatbots were taken from the introduction or conclusions of the article. Sometimes, numerical indicators were used, which were sometimes interpreted by the chatbot not very closely to the text. In general, above-average objectivity scores are Claude (86.8%), Gemini (73.5%), ChatGPT (73%), Copilot (69%), and Perplexity (61.6%). It can be concluded that Claude is best and closest to the sources, conveying their content – that is, it gives the most accurate analysis. However, Gemini (73.5%) and ChatGPT (73%) analyses can also be trusted.

Ethicality. Regarding ethics, as mentioned above, AI chats warn about possible limitations and inaccuracies, but not in their responses. Sometimes Copilot issues such warnings.

Additional parameters. An essential advantage of ChatGPT is its provision of information on the number of words, average handle time, and the number of searches for the query. The most full and informative responses are provided by Gemini (2942 words) and Perplexity AI (2455 words).

5. CONCLUSIONS AND PROSPECTS FOR FURTHER RESEARCH

1. The analysis of scientific publications demonstrates a fundamental transformation of literature review methodology towards a model of 'hybrid intelligence', where the researcher's role evolves from an 'information processor' to an 'architect' of search strategies. Studies confirm the high potential of specialised AI tools (such as Elicit and Scite) for screening and data extraction, enabling the creation of 'living systematic reviews'. Conversely, a 'performance paradox' is evident: universal generative chatbots prove unreliable for autonomous bibliographic searches due to the high risk of 'hallucinations' (fabrication of sources) and lack of capacity for deep conceptual synthesis. Consequently, the effective deployment of AI is viable only within a 'human-in-the-loop' framework, necessitating mandatory expert verification of generated content and transparent documentation of algorithmic interactions to ensure academic integrity.

2. Based on the assessment of 5 AI chatbots (ChatGPT, Copilot, Gemini, Claude, Perplexity AI) conducted by the authors-experts against 8 basic criteria set by the authors and an additional parameter (number of words), the following distinctive features of AI chatbots in terms of searching and analysing research sources were defined:

(1) **ChatGPT** is the most balanced model, which demonstrated 100% accuracy in searching sources, high objectivity (73%), an average relevancy, partial correspondence, quite good structuredness and depth of analysis, but not enough completeness. ChatGPT can be used as a tool for deep searching on a specific topic, which will help analyse the sources found. However, it does not cover many sources and has a strong tendency to interpretation, which reduces the certainty of the answers.

(2) **MS Copilot** generates answers quickly and provides the most relevant answers to the query's topic compared to other AI chatbots studied (score 1.8). It has a high relevance indicator (50.2%), but lower completeness and objectivity indicators (69%). Copilot offers a clear thematic focus and effective search, and it aims to respond to a specific research query rather than collect many sources, which is essential for targeted scientific search.

(3) **Gemini** provides the most volumetric response. It has sufficient depth and objectivity, but its relevance is lower compared to Copilot and Claude. Also, Gemini can be too wordy, moving away from the topic of the query. Gemini performs a detailed collection and generalisation, with lower search accuracy, but is well-suited for a comprehensive literature review.

(4) **Claude** has the highest relevance (56.1%) and objectivity (86.8%), due to the small number of sources used. However, it has the fewest in-text citations (some parts of the text have no citations at all). Claude's responses lack depth, generalisation, conclusions, definitions of tendencies, etc. Claude reproduces source content well and interprets it correctly. This shows the functions of accurate analysis and careful work with the text.

(5) **Perplexity AI** can be considered a leader in source quality (the most significant number of sources from diverse scientific platforms and in multiple languages). Perplexity AI already draws generalisations and conclusions quite well, but it has the least relevance and

objectivity of all the analysed AI chatbots. Perplexity AI functions as a search engine with an extensive collection of sources, but it does not filter them effectively based on content.

3. Based on the analysis of the above-mentioned features of ChatGPT, MS Copilot, Gemini, Claude, and Perplexity AI for literature search and analysis, it is possible to conclude whether AI chatbots can be reliable tools for searching and analysing scientific literature. Trust in the credibility of sources is high (100%), so AI chatbots can be used reliably as a literature search tool. The structuring of AI chatbot responses enables the researcher to navigate the text quickly, select relevant or irrelevant information, organise the literature, and highlight topics. The variability of relevance and objectivity indicators requires the researcher to verify texts generated by AI chatbots and to critically assess the results provided by the AI chatbots for ethical reasons. Thus, the analysed AI chatbots can be used for the initial stage of finding and selecting sources (Perplexity AI is best for this) for scientific articles, for quick familiarisation with the thematic literature, for preliminary analysis, for structured reviews, and for the formation of ideas and the direction of research. However, the researcher cannot wholly rely on the relevance of the sources (some of them may not correspond to the topic), completeness (most AI chatbots, except Perplexity AI, have limited access to scientific databases and publications), analysis of the essence of the article, and ethicality. So, AI chatbots are effective as an initial tool for searching and analysing scientific literature, but require researcher validation to ensure content accuracy, source selection, correctness of analysis, completeness of review, generalisations, and critical analysis.

The results of this study provide a scientific basis for the safe transition of research practices to the 'hybrid intelligence' paradigm. An understanding of the features, advantages and disadvantages of different AI chatbots is necessary for literature analysis, which directly affects the reliability and pace of future scientific discoveries. The identified features of AI chatbots provide a basis for updating university curricula, developing courses in digital literacy, and revising institutional policies on AI use.

Limitations. The article's limitations include the human factor that could have influenced the experts' assessment of AI chatbots' responses. After all, criteria such as relevance and objectivity were assessed by experts based on their experience, which can introduce a degree of subjectivity. The limited scope of the scientific article prevented the authors from comparing AI chatbot results with those from services such as Scopus AI, LeapSpace, Google Scholar Labs, or Web of Science Research Assistant. No comparisons were made between the ordinary (standard) response mode and the deep research mode, nor between searches using the Scopus and Web of Science databases. All this can be the subject of future research. It should also be noted that the results may not be exactly replicated even with the same AI chatbot queries used in the study. This phenomenon is due to the rapid development of technology and, as in the case of Perplexity AI, to the random selection of a model based on the complexity of the query.

Conflict of Interest. Co-author Kateryna Osadcha is a member of this journal's editorial board. She did not participate in the editorial review of the manuscript or in the decision to publish it. The manuscript underwent independent peer review by external experts.

REFERENCES (TRANSLATED AND TRANSLITERATED)

[1] E. Navajas Cawood, M. Vespe, A. Kotsev, and R. Van Bavel, *Generative AI Outlook Report: Exploring the Intersection of Technology, Society and Policy*. Luxembourg: Publications Office of the European Union, 2025, doi: 10.2760/1109679. (in English)

[2] OECD, *OECD Digital Education Outlook 2023: Towards an Effective Digital Education Ecosystem*. Paris, France: OECD Publishing, 2023, doi: 10.1787/c74f03de-en. (in English)

[3] UNESCO, *Guidance for Generative AI in Education and Research*. Paris, France: UNESCO eBooks, 2023, doi: 10.54675/ewzm9535. (in English)

[4] R. Borah, A. W. Brown, P. L. Capers, and K. A. Kaiser, "Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry," *BMJ Open*, vol. 7, no. 2, 2017, Art. no. e012545, doi: 10.1136/bmjopen-2016-012545. (in English)
[5] G. Wagner, R. Lukyanenko, and G. Paré, "Artificial intelligence and the conduct of literature reviews," *J. Inf. Technol.*, vol. 37, no. 2, pp. 209–226, 2022, doi: 10.1177/02683962211048201. (in English)
[6] E.-M. Schön, J. Kollmorgen, M. Neumann, and M. Rauschenberger, "A case study on using generative AI in literature reviews: Use cases, benefits, and challenges," in *Proc. 21st Int. Conf. Web Information Systems and Technologies (WEBIST)*, 2025, pp. 533–543, doi: 10.5220/0013672700003985. (in English)
[7] Y. Li *et al.*, "Enhancing systematic literature reviews with generative artificial intelligence: Development, applications, and performance evaluation," *J. Amer. Med. Inform. Assoc.*, vol. 32, no. 4, pp. 616–625, 2025, doi: 10.1093/jamia/ocaf030. (in English)
[8] F. Bolaños, A. Salatino, F. Osborne, and E. Motta, "Artificial intelligence for literature reviews: Opportunities and challenges," arXiv preprint arXiv:2402.08565, 2024, doi: 10.48550/arXiv.2402.08565. (in English)
[9] B. Burger, D. K. Kanbach, S. Kraus, M. Breier, and V. Corvello, "On the use of AI-based tools like ChatGPT to support management research," *Eur. J. Innov. Manage.*, vol. 26, no. 7, pp. 233–241, 2023, doi: 10.1108/EJIM-02-2023-0156. (in English)
[10] M. Rahman, H. J. R. Terano, N. Rahman, A. Salamzadeh, and S. Rahaman, "ChatGPT and academic research: A review and recommendations based on practical examples," *J. Educ., Manage. Develop. Stud.*, vol. 3, no. 1, pp. 1–12, 2023, doi: 10.52631/jemds.v3i1.175. (in English)
[11] M. Alshater, "Exploring the role of artificial intelligence in enhancing academic performance: A case study of ChatGPT," SSRN working paper, 2022, doi: 10.2139/ssrn.4312358.. (in English)
[12] Y. N. Gwon *et al.*, "The use of generative AI for scientific literature searches for systematic reviews: ChatGPT and Microsoft Bing AI performance evaluation," *JMIR Med. Inform.*, vol. 12, 2024, Art. no. e51187, doi: 10.2196/51187. (in English)
[13] M. Sparkman and A. Witt, "Claude AI and Literature Reviews: An Experiment in Utility and Ethical use," *Library Trends*, vol. 73, no. 3, pp. 355–380, Feb. 2025, doi: 10.1353/lib.2025.a961199. (in English)
[14] J. Miao, C. Thongprayoon, S. Suppadungsuk, O. A. Garcia Valencia, F. Qureshi, and W. Cheungpasitporn, "Ethical dilemmas in using AI for academic writing and an example framework for peer review in nephrology academia: A narrative review," *Clin. Pract.*, vol. 14, no. 1, pp. 89–105, 2024, doi: 10.3390/clinpract14010008. (in English)
[15] M. S. Ya'u and M. S. Mohammed, "AI-Assisted Writing and Academic Literacy: Investigating the dual impact of language models on writing proficiency and ethical concerns in Nigerian higher education," *International Journal of Education and Literacy Studies*, vol. 13, no. 2, pp. 593–604, Apr. 2025, doi: 10.7575/aiac.ijels.v.13n.2p.593. (in English)
[16] C. Zielinski *et al.*, "Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications," *Curr. Med. Res. Opin.*, vol. 40, no. 1, pp. 11–13, 2024, doi: 10.1080/03007995.2023.2286102. (in English)
[17] E. S. Rentier, "To use or not to use: Exploring the ethical implications of using generative AI in academic writing," *AI Ethics*, vol. 5, pp. 3421–3425, 2025, doi: 10.1007/s43681-024-00649-6. (in English)
[18] E. N. Rabbianty, S. Azizah, and N. K. Virdyna, "AI in academic writing: Assessing current usage and future implications," *INSANIA: J. Pemikiran Alternatif Kependidikan*, vol. 28, no. 1a, pp. 14–35, 2023, doi: 10.24090/insania.v28i1a.9278. (in English)
[19] H. Li and X. Wu, "The use of generative AI tools in academic writing: A systematic review of research trends and thematic insights," *AI Ethics*, vol. 5, pp. 5821–5840, 2025, doi: 10.1007/s43681-025-00827-0. (in English)
[20] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008, doi: 10.1017/CBO9780511809071. (in English)
[21] E. Purssell and N. McCrae, *How to Perform a Systematic Literature Review: A Guide for Healthcare Researchers, Practitioners and Students*. 2024. doi: 10.1007/978-3-031-71159-6. (in English)
[22] L. Elder and R. Paul, *Critical thinking: Tools for Taking Charge of Your Learning and Your Life*. Bloomsbury Publishing PLC, 2020. (in English)
[23] C. K. Reddy and P. Shojaee, "Towards scientific discovery with generative AI: Progress, opportunities, and challenges," in *Proc. 39th AAAI Conf. Artif. Intell.*, vol. 39, 2025, pp. 28601–28609, doi: 10.1609/aaai.v39i27.35084. (in English)
[24] L. Kugler, "How do you measure AI?" *Commun. ACM*, vol. 68, no. 4, pp. 15–17, Apr. 2025, doi: 10.1145/3708972. (in English)
[25] N. Shumeiko, K. Osadcha, and M. Spišiaková, *Why artificial intelligence? AI Tools in Information Technology and Foreign Language Education at the Tertiary Level*. Berlin, Germany: Peter Lang Verlag, 2026, doi: 10.3726/b23215. (in English)

[26] UNESCO, "Recommendation on the ethics of artificial intelligence," UNESCO, Paris, France, 2022. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000381137. (in English)
[27] S. Porsdam Mann *et al.*, "Guidelines for ethical use and acknowledgement of large language models in academic writing," *Nat. Mach. Intell.*, vol. 6, pp. 1272–1274, 2024, doi: 10.1038/s42256-024-00922-7. (in English)
[28] European Commission, "Ethics guidelines for trustworthy AI," Digital Strategy, Apr. 08, 2019. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. (in English)
[29] E. Bailyn, "Top generative AI chatbots by market share – December 2025." First Page Sage. https://firstpagesage.com/reports/top-generative-ai-chatbots (accessed Feb. 18, 2026). (in English)
[30] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet Things Cyber-Phys. Syst.*, vol. 3, pp. 121–154, 2023, doi: 10.1016/j.iotcps.2023.04.003. (in English)
[31] T. Wang and Q. Zhu, "ChatGPT – Technical research model, capability analysis, and application prospects," in *Proc. IEEE 7th Adv. Inf. Technol., Electron. Automat. Control Conf. (IAEAC)*, Chongqing, China, 2024, pp. 787–796, doi: 10.1109/IAEAC59436.2024.10503756. (in English)
[32] S. A. Lehr, A. Caliskan, S. Liyanage, and M. R. Banaji, "ChatGPT as research scientist: Probing GPT's capabilities as a research librarian, research ethicist, data generator, and data predictor," *Proc. Nat. Acad. Sci.*, vol. 121, no. 35, 2024, Art. no. e2404328121. (in English)
[33] Microsoft Tech Community, "Reasoning models in Microsoft Copilot: Who's doing the thinking?" https://techcommunity.microsoft.com/discussions/microsoft365copilot/reasoning-models-in-microsoft-copilot-who%E2%80%99s-doing-the-thinking/4429215 (accessed Feb. 18, 2026). (in English)
[34] Microsoft, "Copilot keeps getting better." Microsoft.com. https://www.microsoft.com/en-nz/microsoft-copilot/for-individuals?form=MY02P9 (accessed Feb. 18, 2026). (in English)
[35] K. Osadcha, V. Osadchyi, V. Proshkin, and O. Spirin, "Studying IT educators' satisfaction with using Microsoft Copilot chat to perform professional tasks," *Inf. Technol. Learn. Tools*, vol. 107, no. 3, pp. 153–167, 2025, doi: 10.33407/itlt.v107i3.6184. (in English)
[36] N. Shazeer, "Fast transformer decoding: One write-head is all you need," arXiv preprint arXiv:1911.02150, 2019, doi: 10.48550/arXiv.1911.02150. (in English)
[37] E. Supriyadi, "Exploring Google Bard's (Gemini) role in enhancing research articles in computational thinking and mathematics education," *Papanda J. Math. Sci. Res.*, vol. 3, no. 1, pp. 28–37, 2024, doi: 10.56916/pjmsr.v3i1.707. (in English)
[38] Anthropic, "What's new in Claude 4.5." Anthropic.com. https://platform.claude.com/docs/en/about-claude/models/whats-new-claude-4-5 (accessed Jan. 01, 2026). (in English)
[39] Anthropic, "System card: Claude Opus 4.5," 2025. Anthropic.com. https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf (accessed Jan. 01, 2026). (in English)
[40] Perplexity, "Getting started. What is Perplexity?" Perplexity.ai. https://www.perplexity.ai/hub/getting-started (accessed Jan. 01, 2026). (in English)
[41] Perplexity, "What advanced AI models are included in my subscription?" Perplexity.ai. https://www.perplexity.ai/help-center/en/articles/10354919-what-advanced-ai-models-are-included-in-my-subscription (accessed Jan. 01, 2026). (in English)
[42] Perplexity, "Academic filter guide." Perplexity.ai. https://docs.perplexity.ai/guides/academic-filter-guide (accessed Jan. 01, 2026). (in English)
[43] J. Khanifar, "Evaluating AI-generated responses from different chatbots to soil science-related questions," *Soil Adv.*, vol. 3, 2025, Art. no. 100034, doi: 10.1016/j.soilad.2025.100034. (in English)
[44] M. F. Şahin *et al.*, "Responses of five different artificial intelligence chatbots to the top searched queries about erectile dysfunction: A comparative analysis," *J. Med. Syst.*, vol. 48, 2024, Art. no. 38, doi: 10.1007/s10916-024-02056-0. (in English)
[45] E. Bostan, M. T. Uçar, and E. Dönmez, "Evaluation of the accuracy, reliability, quality, and readability of artificial intelligence chatbots-generated responses to acne-related questions," *Turk. J. Dermatol.*, vol. 19, no. 4, pp. 235–243, 2025, doi: 10.4274/tjd.galenos.2025.06977. (in English)
[46] Perplexity, "What is research mode?" Perplexity.ai. https://www.perplexity.ai/help-center/en/articles/10738684-what-is-research-mode (accessed Jan. 01, 2026). (in English)
[47] K. Osadcha and V. Osadchyi, "Dataset of AI Chatbot Responses to Four Research Questions". Zenodo, May 05, 2026. doi: 10.5281/zenodo.20038076. (in English)

Text of the article was accepted by Editorial Team 11.03.2026.

ШІ ЧАТ-БОТИ В ДОСЛІДНИЦЬКІЙ ПРАКТИЦІ: МОЖЛИВОСТІ ДЛЯ ПОШУКУ ТА АНАЛІЗУ НАУКОВИХ ДЖЕРЕЛ

Катерина Осадча
доктор педагогічних наук, професор, постдокторант
Вільний університет Брюсселя, м.Брюссель, Бельгія
провідний науковий співробітник, Інститут цифровізації освіти НАПН України, м. Київ, Україна
ORCID ID 0000-0003-0653-6423
Kateryna.Osadcha@vub.be, k.osadcha@iitlt.gov.ua

Марина Осадча
магістр освіти, викладач
Мелітопольський державний педагогічний університет імені Богдана Хмельницького,
м. Запоріжжя, Україна
Старша школа Міто Кірьо, Міто, Ібаракі, Японія
ORCID ID 0009-0007-6750-1853
niketa007@gmail.com

Наталія Шумейко,
кандидат педагогічних наук, викладач
Братиславський університет економіки та бізнесу, м. Братислава, Словацька Республіка
ORCID ID 0000-0002-7859-7519
natalia.shumeiko@euba.sk

Володимир Прошкін
доктор педагогічних наук, професор, викладач математики
Ротеремський коледж, м. Ротерем, Великобританія
ORCID ID 0000-0002-9785-0612
Volodymyr.Proshkin@rnngroup.ac.uk

Анотація. У сучасній дослідницькій практиці ШІ-чати отримують визнання як ефективні цифрові помічники, що полегшують пошук та аналіз наукової літератури. У цьому контексті у статті представлено комплексний аналіз особливостей використання ШІ-чатів у дослідницькій практиці, зокрема для пошуку та аналізу наукових джерел. Методологія дослідження базується на змішаному підході, який поєднує аналіз наукових джерел з порівняльною оцінкою результатів 5 ШІ-чатів (ChatGPT, Microsoft Copilot, Gemini, Claude, Perplexity AI) за 8 критеріями. Для аналізу експертами з чотирьох галузей знань (інженерія, природничі науки, гуманітарні науки, соціальні науки) використовувались кількісні та якісні метрики оцінювання відповідей ШІ-чатів на однакові дослідницькі запити. Порівняльний аналіз ШІ-чатів показав їхні різні можливості та обмеження в роботі з науковими джерелами. ChatGPT є найбільш збалансованим інструментом, з високою точністю та об'єктивністю відповідей, але він не охоплює багато джерел та схильний до інтерпретацій. MS Copilot відрізняється швидкістю генерації відповідей та достатньою релевантністю джерел, але поступається в об'єктивності аналізу. Gemini забезпечує достатню відповідність темі та об'єктивність у відповідях, але відповіді часто надмірно довгі та менш точні. Claude демонструє найвищі показники релевантності та об'єктивності працюючи з невеликою кількістю джерел, але майже не робить узагальнень. Perplexity AI є лідером за обсягом та якістю джерельної бази, добре формує узагальнення, але характеризується найнижчою релевантністю джерел та об'єктивністю відповідей. Результати показали, що всі досліджені ШІ-чати демонструють високий рівень формальної надійності джерел, але суттєво відрізняються за релевантністю джерел, аналітичною глибиною та здатністю адекватно інтерпретувати науковий зміст статей. ШІ-чати є ефективними інструментами для початкової орієнтації в темі дослідження, швидкого збору джерел для аналізу та попереднього структурування матеріалу. Однак вони не можуть повноцінно провести глибокий та критичний аналіз наукових джерел з конкретного дослідницького питання.

Ключові слова: штучний інтелект; чат-боти зі штучним інтелектом; дослідницька практика; оцінка інструментів штучного інтелекту; пошук літературних джерел за допомогою штучного інтелекту; аналіз академічних джерел за допомогою штучного інтелекту.

This work is licensed under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.