Wagner G., Prester, J. Mousavi, R., Lukyanenko R., and Paré G. (2026). Generative Artificial intelligence for literature reviews, ***Journal of Information Technology***. *Open access*

Generative Artificial Intelligence for Literature Reviews

Gerit Wagner
University of Bamberg

Julian Prester
University of Sydney

Reza Mousavi
University of Virginia

Roman Lukyanenko
University of Virginia

Guy Paré
HEC Montréal
guy.pare@hec.ca

Abstract. Generative artificial intelligence (GenAI), based on large-language models (LLMs), such as ChatGPT, has taken organizations, academia, and the public by storm. In particular, impressive GenAI capabilities such as summarization of large text corpora, question-answering, data extraction, and translation, carry profound implications for the conduct of literature reviews. This impacts science, organizations and the general public, as all can benefit from GenAI-supported literature reviews. Building on the technical foundations of GenAI and grounded in established methodological discourse, this work outlines approaches for conducting literature reviews using both general-purpose (e.g., ChatGPT, Gemini, Claude) and specialized GenAI tools (e.g., Consensus, Elicit). We provide illustrative examples of prompts and suggest methodologically-sound literature review strategies. Throughout this perspective paper, we adopt a balanced approach considering both the opportunities and the risks of relying on GenAI in the conduct of literature reviews. We conclude by discussing philosophical questions related to the effects of GenAI on long-term scientific progress, and also present fruitful opportunities for research on improving the core of GenAI's technology – its architecture and training data - and suggest open issues in GenAI-based literature reviews methodology.

Keywords: Literature review, generative artificial intelligence, GenAI, large language models, LLMs, ChatGPT, Claude, Gemini

Generative Artificial Intelligence for Literature Reviews

Introduction

Generative artificial intelligence (GenAI), particularly in the form of large language models (LLMs) such as ChatGPT, has rapidly gained visibility across organizations, academia, and public discourse. It is widely viewed as a potentially transformative development, especially for knowledge-intensive tasks involving language, including summarization, question answering, and synthesis. At the same time, assessments of GenAI's impact remain uneven and contested. While some users report substantial gains in efficiency and convenience, others point to disappointing performance, limited real-world value realization, and recurring cycles of hype and disillusionment associated with earlier waves of AI adoption.

Notwithstanding these divergent views, the pace of recent advances in GenAI has been striking, raising fundamental questions about how such systems may reshape established scientific practices. Although GenAI holds considerable promise for advancing research, it also presents significant challenges. The exponential growth of scientific publications already imposes a substantial cognitive load on researchers, increasing the likelihood that they overlook relevant and timely findings (Bornmann et al. 2021; Thelwall and Pardeep 2022). In this context, GenAI introduces a profound paradox: while these technologies may further accelerate the production of scholarly content, thereby intensifying informational overload, they also offer powerful new capabilities for synthesizing large bodies of literature and mitigating the very complexity they help create.

Within the specific domain of literature reviews, these affordances can support a wide range of activities, from foundational tasks such as exploratory searching and summarization to more complex functions involving project management and conceptual knowledge synthesis (Alavi et al. 2024; Schryen et al. 2024; Susarla et al. 2023). However, not all observers are convinced of

their unqualified benefits. Some caution that an overreliance on GenAI may erode core scientific skills by discouraging deep engagement with primary sources (Zur Schlemmer 2024). Accordingly, while GenAI may facilitate more comprehensive and efficient evidence synthesis, its role in scientific inquiry warrants careful and balanced consideration of both its potential benefits and its associated risks.

This need for caution is reinforced by the considerable uncertainty surrounding GenAI's ongoing development. It thus remains unclear how GenAI will ultimately reshape research, as the broader GenAI ecosystem - including transformer architectures, pre-training pipelines, alignment and safety protocols, and the applications built on them - continues to evolve rapidly. Compounding this uncertainty is the continual discovery of emergent properties and unforeseen functionalities within GenAI—a phenomenon whereby novel, qualitatively distinct capabilities, such as multi-step reasoning, manifest only after models surpass a critical threshold of scale (Wei et al., 2022a; Zoph et al. 2022). Consequently, many advanced capabilities are not the product of targeted engineering but are instead outcomes of revised scaling principles and empirical discovery (Hoffmann et al., 2022; Kaddour et al. 2023).

This discovery-oriented paradigm poses fundamental challenges to interpretability, leaving significant gaps in understanding model mechanisms and in reliably steering their behavior (Bowman, 2023). The continued adaptation of these systems to novel problems and domains further underscores how little is known about the ultimate boundaries of their capabilities. Unsurprisingly, experts hold widely diverging expectations for the future of GenAI, ranging from its containment within narrow, regulated use cases to the eventual emergence of artificial general intelligence (Bubeck et al., 2023; Hubert et al., 2024).

Considering the existing uncertainty related to the future of GenAI, we believe it is instructive to discuss possible opportunities, modalities, and risks related to the use of GenAI in the conduct of

literature reviews. While the initial discourse has quickly produced suggestions on how GenAI could be of use in our context (e.g., Alshami et al. 2023; Rahman et al. 2023; Temsah et al. 2023), these preprints, commentaries, and studies do not offer a substantial connection to the established methodological knowledge. Without considering how GenAI-enhanced literature reviews can be reconciled with the goals and types of reviews (Paré et al. 2015), the activities of the process, systematicity of methodological choices, as well as reporting requirements (Paré et al. 2016; Templier and Paré 2018), the use of GenAI for literature reviews may struggle to produce reviews of good quality and fail to meet expectations of an accepted and valid review process. Instead, it remains essential that researchers are knowledgeable in their domain, understand the nuances of literature review methods, and leverage GenAI considerately (Qureshi et al. 2023). In doing so, we believe it is important to discuss how established methodological practices should be continued and how GenAI can enhance the conduct of literature reviews.

The primary goal of this paper is to discuss how GenAI transforms the conduct of literature reviews, provide constructive suggestions for prospective authors, and discuss potential risks and opportunities. It continues our work on the use of previous generations of AI[1] for literature reviews (Wagner et al. 2022) and discusses how recent advances in GenAI could affect literature review practices in the future.

[1] The approaches considered by Wagner et al. (2022) included popular AI techniques such as neural networks, random forests, decision trees, and pre-transformer natural language processing algorithms, such as LSA or LDA. These AI techniques remain popular and valuable for literature reviews, as outlined in the aforementioned paper. However, GenAI offers new opportunities and challenges not covered by and not relevant to these techniques (such as hallucinations and specific biases).

This paper is relevant for literature reviews across a wide range of scientific disciplines and research genres. It is also written for a broader audience, as organizations, journalists and the public increasingly use GenAI to conduct reviews of their own. At the same time, we deliberately choose a specific context for our examples. Our running examples are GenAI-supported reviews in the context of information systems design and use. First, information systems design and use is the context we are most familiar with. We have undertaken numerous manual reviews and reviews with the support of previous generations of AI tools, examining various aspects of information systems design and use (Dissanayake et al. 2025; Larsen et al. 2025; Prester et al. 2021; Recker et al. 2021). We are well positioned to interpret the performance and findings of GenAI tools in this context.

Furthermore, information systems, as a proximal discipline to GenAI, is already actively involved in understanding the use of GenAI (e.g., Alavi et al., 2024; Ngwenyama and Rowe, 2024, Storey et al., 2025; Susarla et al., 2023). Finally, the information systems discipline has developed a vibrant methodological discourse on literature reviews, including on AI-supported reviews (e.g., Boell and Cecez-Kecmanovic 2015; Paré et al. 2024; Storey et al. 2025; Templier and Paré 2018; Wagner et al. 2022), increasingly serving as a reference to other disciplines (e.g., Aguinis et al. 2023).

The remainder of this paper is structured as follows. First, we establish a conceptual foundation in GenAI, distinguishing it from traditional AI, and surveying its primary modalities. We also introduce effective prompting strategies as the core method for interacting with these systems. Next, we build on these concepts to offer suggestions for the use of GenAI and corresponding prompts for the different activities of the review process. Finally, and before concluding the paper, we discuss the broader opportunities, challenges and open questions related to the use of GenAI in the conduct of literature reviews.

Technological Foundations

GenAI vs. AI

GenAI represents a significant evolution from traditional AI, shifting the technological paradigm from data analysis to content creation. Traditional AI approaches are primarily discriminative; they excel at pattern recognition, classification, and predictive analytics based on existing datasets. Their main purpose is to interpret and make judgments about patterns in data. In contrast, GenAI models are different: their core function is to produce new, synthetic content that mimics the patterns and structures of the data on which they were trained. This capability spans a wide array of modalities, allowing these systems to compose text, create realistic images, write computer code, synthesize audio, and even design molecular structures (Brown et al. 2020; Zhong et al. 2024). This marks a fundamental shift from technologies that primarily analyze existing information to those that can synthesize novel artifacts.

This distinction is profound in the context of research. While a traditional model might classify scholarly articles or predict trends, GenAI can actively participate in the research process. For instance, it can assist in problem formulation by drafting hypotheses (text) or creating conceptual diagrams (images). It can enhance data analysis by generating code to process datasets or by creating synthetic data for model validation. For dissemination, it can draft manuscript sections (text), design figures (images), or even compose a score for a video abstract (audio). This shift from merely automating repetitive tasks to actively contributing creative input and generating new content underscores the transformative potential of GenAI in redefining the conduct of AI-supported literature reviews (AILRs).

GenAI Modalities

The diverse applications outlined above are enabled by distinct classes of generative models, each specialized for a specific data modality. The primary modalities include text, image, audio, video, and integrated multimodal systems.

The text modality is foundational to GenAI, powered by the immense capabilities of LLMs such as GPT-4 and Llama 3.1. The development of these models marks a pivotal departure from earlier deep learning approaches in Natural Language Processing (NLP). For years, the field was dominated by sequential architectures like Recurrent Neural Networks (RNNs) and their more advanced variant, Long Short-Term Memory (LSTMs). These models processed text one word at a time, maintaining a "memory" of prior context. However, this sequential method faced two critical limitations: first, it struggled to maintain context over long passages, as information would degrade or "vanish" across many steps (Zhao et al. 2020); second, its word-by-word nature prevented the parallelization needed to train on internet-scale datasets (Hwang and Sung 2015).

The introduction of the Transformer architecture in 2017 was a paradigm shift that solved these problems (Vaswani et al. 2017). Its key innovation, the self-attention mechanism, abandoned sequential processing entirely. Instead, it allowed the model to weigh the influence of all words in a sequence simultaneously, creating direct pathways for context to flow regardless of distance and thus capturing complex, long-range dependencies.

Crucially, this architecture's design was highly parallelizable, making it computationally feasible to train models of unprecedented size (Devlin et al. 2019). This scalability is what directly paved the way for modern LLMs (Radford et al. 2018). By dramatically increasing model parameters and training data, researchers discovered that scaled-up Transformer models exhibited the sophisticated, emergent capabilities that have since catalyzed a revolution across NLP, enabling applications from nuanced conversational agents to complex code generation (Brown et al. 2020).

Concurrently with the revolution in text generation, a separate lineage of architectural innovation was enabling the synthesis of rich media. Beginning with Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and later advancing significantly with diffusion models (Ho et al., 2020; Sohl-Dickstein et al., 2015), these techniques resulted in models that generate high-fidelity images, audio, and video from descriptive text prompts. This progress is exemplified by tools like DALL·E[2] for image generation and Sora[3] for video generation, which can translate linguistic concepts into detailed visual content.

The current frontier of GenAI is defined by the convergence of two powerful streams: the linguistic prowess of Transformer-based LLMs and the sensory generation capabilities of architectures like diffusion models. The evolution toward today's multimodal systems began with foundational techniques designed to bridge the gap between different data types. A pioneering step in this direction was the development of joint embedding spaces, exemplified by models like CLIP (Contrastive Language–Image Pre-training) (Radford et al., 2021). By learning to align text and images within a shared representational framework, CLIP enabled models to achieve a cross-modal understanding for tasks like zero-shot image classification and text-based image retrieval, laying the essential groundwork for more complex integrations.

The evolution from these initial integrations to today's state-of-the-art models has been rapid, marked by a fundamental shift in architectural philosophy. Whereas early multimodal applications often relied on a pipeline of separate, specialized models (e.g., a speech-to-text model feeding into an LLM, which then outputs to a text-to-speech model), the current frontier is

[2] https://openai.com/index/dall-e-3/

[3] https://openai.com/sora/

defined by single, unified models trained end-to-end. This paradigm shift is exemplified by Google's Gemini, which was designed to be "natively multimodal" from its inception, capable of reasoning seamlessly across text, images, video, and audio within one cohesive architecture (Gemini Team Google 2023), resulting in tools such as NotebookLM[4]. Similarly, OpenAI's GPT-4o ("o" for "omni")[5] replaced its prior model pipeline with a unified system, drastically reducing latency and enabling fluid, real-time interaction across text, audio, and images. This architectural leap allows current models to perform highly complex cross-modal tasks, such as answering verbal questions about a live video feed or interpreting emotional tone from audio, within a single, coherent system.

While the ability of these state-of-the-art models to process text, audio, and video represents a transformative frontier for many research fields, their application to the literature review process hinges primarily on their sophisticated textual capabilities. Scholarly knowledge is overwhelmingly codified and disseminated through text, making the core activities of a literature review—from source identification to synthesis and narrative construction—fundamentally text-centric endeavors.

Accordingly, our analysis focuses on GenAI models renowned for their robust text processing. This includes both LLMs like GPT-4 and Llama 3.1[6], and leading multimodal systems such as GPT-4o, Gemini, and Claude 3.5 Sonnet, whose underlying linguistic engines are paramount for

[4] https://notebooklm.google/

[5] https://openai.com/index/hello-gpt-4o/

[6] Detailed explanations of LLM architectures and training paradigms are provided in the supplemental materials available online.

this work.[7] Harnessing the power of these models for academic purposes requires deliberate interaction strategies (though we do explore opportunities for non-textual formats in a later section). The subsequent section, therefore, delves into methodologies for interfacing with GenAI, emphasizing prompting techniques that optimize its effectiveness in research settings.

Prompting Strategies for GenAI

Given the intrinsic dependence of LLMs and multimodal GenAI on the initial prompt for generating text, the selection of an appropriate prompt is critical. A spectrum of prompting strategies exists, each tailored to enhance the model's performance in specific contexts. In the subsequent discussion, we delve into various prompting strategies that hold particular relevance for conducting literature reviews. These strategies are designed not only to refine the model's output in terms of relevance and specificity but also to ensure that the synthesized reviews are comprehensive, accurately reflecting the breadth and depth of the existing scholarly discourse. This approach underscores the necessity of strategic prompt design as a fundamental step in leveraging GenAI for academic and research purposes, particularly in the meticulous task of a literature review. When applying GenAI to literature reviews in academic research, several prompting strategies stand out for their effectiveness:

1. Exploratory prompting: This strategy involves asking open-ended questions to explore broad themes or identify under-researched areas within a field (Sun and Wang 2025). This approach mirrors exploratory search and information-seeking behaviors that researchers employ when navigating unfamiliar domains or scoping new research

[7] Please note that multimodal GenAI such as GPT-4o outperform text-only LLMs such as GPT-4 even in natively text-based tasks. Please refer to https://crfm.stanford.edu/helm/lite/latest/ for a benchmark of multimodal GenAI models and LLMs.

directions (Gusenbauer and Haddaway 2021). For example, it is particularly effective in the exploratory stages of a literature review, where the goal is to map out the landscape of existing research. An example of the exploratory prompting strategy applied to the literature review process is provided in problem formulation prompts (Tables 1 and 3).

2. Zero-shot and few-shot learning (Touvron et al. 2023): These techniques are particularly useful when dealing with highly specialized or emerging topics in research, where pre-existing examples or detailed training data may be limited. For example, by providing the model with a definition of a task, such as the formulation of a search strategy, and possibly a few examples (few-shot) or none at all (zero-shot), researchers can prompt GenAI to generate insights or identify trends in a literature corpus that have not been explicitly programmed into its training data. An example of the few-shot prompting strategy applied to the literature review process is provided in the search query prompt (Table 5).

3. Chain-of-thought prompting (Wei et al. 2022b): This strategy involves guiding the GenAI model through a logical reasoning process, breaking down complex queries into simpler, sequential steps. For example, when exploring the impact of digital transformation on organizational culture, a researcher might use a chain of thought prompting to first identify key components of digital transformation, then assess their influences on different aspects of organizational culture, and finally synthesize these impacts into a coherent narrative. This step-by-step approach helps in structuring the literature review process and ensures that the model's outputs are not only relevant but also logically sound. An example of the chain-of-thought prompting strategy applied to the literature review process is provided in the data extraction prompt or the data analysis and synthesis prompt (Table 9).

4. Retrieval-augmented generation (RAG) (Lewis et al. 2020): RAG is a technique that enhances LLMs by combining the pre-trained LLM model with additional sources provided by the user before generating a response. This makes the GenAI responses more context-aware, relevant to the specific task, and less prone to "hallucinations." Although RAG is not strictly a prompting strategy, it integrates the generative capabilities of models with the strength of information retrieval, enhancing the quality and relevance of responses. This makes it particularly valuable for complex tasks like literature reviews. For example, in the context of reviewing literature on the impact of digital transformation on organizational culture, a RAG prompt can leverage both the retrieval of existing scholarly articles and the generative aspect to synthesize and analyze findings. An example of the RAG prompting strategy applied to the literature review process is provided in the first literature search prompt (Table 4).

In addition to these strategies, role-based or persona prompting can enhance output quality by prefixing prompts with statements such as 'You are an expert in...' This technique establishes the model's context and expertise level, activating domain-relevant knowledge and improving response quality (Salewski et al. 2023). Several prompts in this paper employ this technique to guide the model toward more specialized outputs.

In sum, by leveraging these prompting strategies properly and responsibly, researchers can harness the capabilities of GenAI to conduct more efficient, thorough, and insightful literature reviews. These approaches not only save time but also enhance the depth and breadth of the review process, enabling scholars to uncover novel insights and contribute more meaningfully to their fields of study.

Applications of GenAI in the Literature Review Process

We now provide the most prevalent opportunities for applying GenAI in the conduct of literature reviews, to sensitize readers to methodological nuances when using GenAI, and to anticipate how GenAI may change the review process in the future. To accomplish this, we adopt an iterative conception of the literature review process, in which researchers select and revisit literature review activities without following a strict, predefined sequence (cf. Boell and Cecez-Kecmanovic 2014). For each review activity, we provide example prompts and short evaluations based on the authors' assessment of prompt outputs.

In discussing GenAI capabilities, we refer to HuggingFace[8], an online platform that provides a toolkit library, open-source LLMs, an active community, and educational resources for deep learning-based models, and HELM[9], which provides an overview of state-of-the-art LLM evaluation results across a variety of tasks. While models covered in HuggingFace can be applied to different types of data (including audio, photos, and video), the NLP and multimodal capabilities are particularly interesting for our purposes. They include text generation, question answering, summarization, translation, document question answering, image-text-to-text, and any-to-any format generation.

There are three key premises for our work. First, researchers must be familiar with common methodological practices, such as goals of reviews (Paré et al. 2015; Rowe 2014; Paré et al. 2023), choices in different steps (Templier and Paré 2018), as well as transparency and systematicity requirements (Paré et al. 2016). At least for now, GenAI needs explicit prompts and

[8] https://huggingface.co/tasks

[9] https://crfm.stanford.edu/helm/lite/latest/

context to provide adequate output. In the future, custom GPT versions may combine instructions with extra knowledge and any combination of skills. Second, we expect researchers to ensure that GenAI tools have access to relevant full-text documents, typically by downloading PDFs and providing them with the prompts. This aligns with the reporting requirements for standalone reviews, which ask authors to control, track and report on the retrieved, screened, and analyzed papers (Templier and Paré, 2018). It is important to note that GenAI tools offered by individual publishers, or operating on open-access papers without considering the specific sample of the review project, are more suitable for informal reviews or exploratory activities, rather than the full standalone review process. Third, researchers must be aware that entering the prompts does not guarantee a useful outcome for all review projects and that adequate oversight is mandatory. This means that all results provided by GenAI must be fact-checked, evaluated critically, and disclosed appropriately. As Tingelhoff et al. (2025) put it, researchers must carefully evaluate "what we should allow Gen.AI to do" (p.78). In addition, it needs to be checked whether full-text documents (PDFs) provided with the prompt fit into the review scope, or whether task-splitting or paid API-access is needed. To support these activities, corresponding research software will need to offer new functionality related to transparent versioning and validation of literature review data.

Problem Formulation

When embarking on a standalone review paper, the problem formulation involves identifying a promising opportunity, assessing the feasibility of the project, and preparing the groundwork for the review (Müller-Bloch and Kranz 2015; Templier and Paré 2018). We expect summarization, language translation, and question-answering capabilities of GenAI to provide useful support in each of these activities. These capabilities can enable teams to develop and assess different options for review projects. Going beyond the informal chartering activities, GenAI can compile evidence from the literature and offer an initial indication of which type of review aligns well

with the current state of research. Risks of replicating existing reviews, possibly due to the use of non-standardized terminology or even due to publication in a different language, can be reduced by dedicated prompting strategies. In addition, once the review objectives are determined, GenAI may be applied to articulate the rationale for the review, position it relative to other review papers, and assemble the conceptual foundations and key definitions.

Table 1. Prompt to identify prior review papers based on citation context

GenAI-Capability	Data extraction
Prompting Strategy	Exploratory prompting
Requirements	LLMs with file upload and large context window (> 100,000 tokens [a])
Prompt Example	**Upload a selection of relevant papers (PDFs)**[10] Considering the in-text citations of each paper, do the papers refer to prior (standalone) literature reviews? Which ones? Consider all PDFs and state explicitly when you encounter problems in extracting text from the PDF document.
Initial Evaluation[11]	
Outcome	Successful identification of 5 out of 8 literature review papers (GPT-4o), 6 out of 8 literature review papers (Claude 3.5 Sonnet), and 5 out of 8 literature review papers (Gemini 1.5 Pro)

[a] Tokens are the basic units of text that LLMs process; tokens can be short words (e.g., "the") or parts of a word (e.g., "play" and "ing" in *playing*).

The example prompts illustrate how GenAI can be used to support the identification of prior review papers (Table 1), conceptual definitions (Table 2), and suitable review types (Table 3). Reasonable responses can be achieved when an initial set of relevant papers is provided as input,

[10] When using PDF documents in a prompt, it is important to check potential restrictions in context windows, and considering options like splitting data input or using paid APIs.

[11] Additional detail on the evaluation is provided in in the supplementary online material. This paper does not aim to provide a comprehensive benchmarking system for evaluating GenAI's performance in conducting literature reviews. However, we share several useful evaluation mechanisms. We also encourage future research to develop robust benchmarking frameworks for this purpose, building on works like Jin et al. (2021), which focused on the medical domain.

taking advantage of GenAI's PDF reading capabilities. In addition, it is advisable to add instructions specifying the expected output when nothing is found and to provide a clear definition of the review types in question. The latter could be done with reference to submission requirements at target journals (Rivard et al. 2018) and the methods discourse (Paré et al. 2015). We note that none of the example prompts can be answered based on titles and abstracts alone; all require an analysis of full-text papers.

Table 2. Prompt to extract concept definitions

GenAI-Capability	Data extraction
Prompting Strategy	Zero-shot prompting
Requirements	LLMs with file upload and large context window (> 100,000 tokens)
Prompt Example	***Upload a selection of relevant papers (PDFs)***[7] From the PDFs provided, extract the definitions for [add description here]. Provide a direct quote if remote work is defined in the paper. If there is none, state that there is no clear definition of [add topic label here]in the PDF.
Initial Evaluation[8]	
Selected Example	Remote work
Outcome	Successful extraction of 3 out of 5 definitions (GPT-4o), 3 out of 5 definitions (Claude 3.5 Sonnet), and 4 out of 5 definitions (Gemini 1.5 Pro)

Table 3. Prompt to assess the fit of a selected review type

GenAI-Capability	Text analysis and recommendations
Prompting Strategy	Exploratory prompting
Requirements	LLMs with file upload and large context window (> 100,000 tokens)
Prompt Example	**Upload a selection of relevant papers (PDFs)** [7] Definition: A "qualitative systematic review" aims at collecting and aggregating empirical evidence from primary studies to test a narrowly defined hypothesis or model. It is suitable for established topics for which the research questions are narrow and the focus is on qualitative or quantitative empirical studies. Assess a review project focusing on [add description here]. Would a qualitative systematic review be a suitable review type? Provide reasons related to the topic maturity, the scope of research questions, and the nature of prior work.
Initial Evaluation[8]	

Selected Example	Future of work
Outcome	Convincing identification and justification of 2 out of 5 review types (GPT-4o), 1 out of 5 review types (Claude 3.5 Sonnet), and 1 out of 5 review types (Gemini 1.5 Pro)

Literature Search

The literature search commonly proceeds from an exploratory to a systematic search phase (Gusenbauer and Haddaway 2021), and, in the context of GenAI, may increasingly intertwine with exploratory skimming and reading activities (Boell and Cecez-Kecmanovic 2014; Palani et al. 2023; Wagner et al. 2020). One of the key challenges in the traditional process is that the massive volumes of research output resulting from literature searches quickly exceed human information processing capacities (Larsen et al. 2019). When researchers have limited prior knowledge of the literature, it is even harder to identify relevant papers and to direct exploratory reading activities. Emergent GenAI capabilities, like summarizing, classifying, and question-answering, may effectively enable researchers to overcome these limits, engage with the contents of larger sets of papers, and gain insights to adjust search activities. As such, we expect that GenAI capabilities can facilitate more pronounced exploratory search activities (Gusenbauer and Haddaway 2021), complement strictly matching search strategies with semantic searches that include synonyms, as well as support the convergence between search and initial screening, skimming, and reading activities.

As an example of exploratory search capabilities, GenAI can be used for question answering or for generating summaries in the form of tables or graphs, providing researchers with a high-level overview of paper contents (Alshami et al. 2023). Table 4 illustrates this type of prompt. Promising online services in this area are often based on publicly available metadata, such as

abstracts and open-access PDFs (e.g., Paperdigest[12] or scite.ai[13]) or even full text-documents of individual publishers (e.g., Scopus AI[14]). Additionally, researchers may take advantage of tabular summaries offered by services like Consensus[15], or Elicit[16]. Providing access to, and requiring explicit connection to research papers enables these services to address the problems of hallucinations or fictitious references more effectively compared to early tests with generic GenAI tools (McGowan et al. 2023).

Table 4. Prompt to explore prior research using a tabular overview

GenAI-Capability	Text summarization
Prompting Strategy	Retrieval-augmented generation (RAG)
Requirements	LLMs with Retrieval Augmented Generation functionality, such as Consensus or Elicit
Prompt Example	How does [add label of variable] affect the relationship between [add antecedent variable or intervention] and [add outcome variable] in the context of [add context description]? Summarize relevant empirical papers with an abstract summary, the research method, and the key findings.
Initial Evaluation[8]	
Selected Example	Effect of skills on the relationship between LLM support and individual productivity in the context of software development
Outcome	From the first 10 papers returned, 2 were relevant and 8 not relevant (Elicit), 2 were relevant, and 8 not relevant (Consensus). All summaries were adequate and almost exclusively based on the abstracts. Concerning the nature of sources, Elicit returned 4 preprints and 2 papers from reputable journals. Consensus returned 2 preprints and 4 papers from reputable journals.

[12] https://www.paperdigest.org/

[13] https://scite.ai/

[14] https://www.elsevier.com/products/scopus/scopus-ai

[15] https://consensus.app/

[16] https://elicit.com

For systematic searches, which typically involve the design of Boolean search queries for academic databases, use of GenAI has major caveats but also promises to alleviate key challenges. Early experience reports repeatedly confirmed problems with hallucinations (McGowan et al. 2023), as well as responses that focus on openly accessible papers published in emergent outlets while missing most of the major contributions in the field. In addition, GenAI tends to lack access to paywalled content, recent publications, and unpublished work[17] potentially containing valuable findings on non-significant relationships. As such, few expect GenAI, such as ChatGPT, to replace the established retrieval process from academic databases in the near future.

Despite the shortcomings, GenAI offers a promising tool to facilitate and improve the design of systematic search strategies for established search infrastructure. Initial work indicates that GenAI performs well in handling (statistical) synonymy associations (Min et al. 2023; Thießen et al. 2023), which is one of the persistent challenges in finding prior research in many social science disciplines (Larsen and Bong 2016). In fact, identifying and grouping synonyms is at the core of constructing Boolean search queries following the building block approach, which refers to "dividing a query into Facets A, B, and C, complete with variants and synonyms, and then adding these concepts together using the Boolean AND operator" (Booth 2008). In addition, general-purpose GenAI and specialized tools (e.g., DeepL) continue to improve in language translation tasks, offering the possibility to add terminology in different languages and with spelling variations to each concept block. Accordingly, initial research has evaluated the effectiveness of ChatGPT for writing Boolean search queries and found that results are

[17] Obtaining access to unpublished work is essential in meta-analyses to address the "file-drawer problem" and reduce publication bias arising from the underrepresentation of non-significant results. Unpublished studies are typically acquired through personal communication, mailing lists, or institutional repositories.

particularly useful for reviews in which highly precise searches are acceptable, such as rapid reviews (Wang et al. 2023). As such, initial queries, such as those returned by the prompt example (Table 5), can already be used as a starting point, with further expansion of search terms needed to achieve adequate recall.

Table 5. Prompt to suggest an initial search query

GenAI-Capability	Content generation
Prompting Strategy	Few-shot prompting
Requirements	LLMs
Prompt Example	You are an information specialist who develops Boolean queries for systematic reviews. You have extensive experience developing highly effective queries for searching the information systems literature. Your specialty is developing queries that retrieve as few irrelevant documents as possible and retrieve all relevant documents for your information needs. You are able to take an information need such as: "Review of IT Business Value" and generate valid Web of Science queries such as: "TI=(IT OR IS OR …) AND TI=(value OR payoff OR …) AND TI=(firm OR business OR …)". Now you have your information need to conduct research on "[add topic here]", please generate a highly effective systematic review Boolean query for the information need.
Initial Evaluation[8]	
Selected Example	Effect of LLM on individual performance at work
Outcome	Best performing prompt (without examples) out of five prompts as evaluated on GPT-3.5 (Wang et al. 2023)

Literature Screening

In the literature screening phase, researchers label papers as relevant or irrelevant to the review, based on metadata or based on full-text documents (Templier and Paré 2018). Given that only limited information (such as titles and abstracts) is available in the first screen, it is a good practice to retain borderline cases for the second screen. In the second stage, final inclusion decisions are made by examining the paper, and by documenting reasons for inclusion or exclusion in the form of screening criteria and reporting descriptive statistics on the screening process, e.g., in line with the PRISMA standard (Page et al. 2021). The prevalent approach to

controlling the reliability of the screening process, is to have two (or more) researchers screen a sample redundantly, and to measure inter-coder reliability.

Initial research has explored the possibility of using GenAI for screening research papers, concluding that classification performance is currently not accurate enough to automate the process (Castillo-Segura et al. 2023; Syriani et al. 2024). Specifically, the work of Syriani et al. (2024) reports how the performance of GPT3.5, using the best performing prompt template displayed in Table 6, compared to the traditional AI classifiers. While the consistency of screening decisions for individual records was relatively high, it is noteworthy that metrics for LLM-based classification vary considerably across datasets, indicating limited generalizability. In particular, the findings show that the essential recall metric does not dominate the performance of random classification in all cases. As such, follow-up research is needed to determine whether larger models or different prompting strategies can improve classification performance. More generally, it is important to remember that language-based AI techniques are not the only ways to automate and facilitate literature reviews. Traditional AI technologies, such as those based on in-house trained neural networks (Wagner et al. 2022) may be a better choice, especially when high accuracy and reliability are needed.

Table 6. Prompt to screen papers based on title and abstract

GenAI-Capability	Text analysis and recommendations
Prompting Strategy	Zero-shot prompting
Requirements	LLMs
Prompt Example	**Extract abstracts locally and provide them with the prompt** [7] Context: I am screening papers for a systematic literature review. The topic of the systematic review is [add topic here]. The study should focus exclusively on this topic. Instruction: Decide if the article should be included or excluded from the systematic review. I give the title and abstract of the article as input. Only answer include or exclude. Be lenient. I prefer including papers by mistake rather than excluding them by mistake. Task i:

	- Title: "Twelve tips to leverage AI for efficient and effective medical question generation" - Abstract: "Crafting quality assessment questions in medical education […]"
Initial Evaluation[8]	
Selected Example	Effect of generative AI on individual productivity for programmers
Outcome	Best performing prompt that maximizes F2 scores as evaluated on GPT-3.5 (Syriani et al. 2024)

Against this background, we expect that GenAI may find a range of applications in support of the screening process. First, in the case of rapid reviews (common in non-scientific outlets), it may be acceptable to trade-off recall against quick completion of the review process, for instance to inform quick and informal (Syriani et al. 2024) and non-mission critical (Lukyanenko et al. 2025). Second, the capabilities of GenAI may be particularly suitable to complete large-scale reviews. As such, we may see examples complementing the work of Larsen et al. (2019) to cover review topics that do not focus on a particular theoretical model with distinct constructs. Third, GenAI can facilitate a range of preparation tasks, including the development of training materials, examples, and process documentation for coders. These can be later shared to increase transparency and replicability (Burton-Jones et al. 2021; Hevner et al. 2024). Fourth, capabilities of translation can be particularly useful in the screening activities to overcome prevalent language and geographical biases (van Wee and Banister 2023), as suggested in Table 7. Fifth, GenAI can be applied to implement publication filters, such as restrictions to empirical studies required in meta-analyses. Given that researchers often use catchy, rather than descriptive titles[18], it may be practical to apply such filters in the full-text screening stages rather than in the search (Higgins et al. 2023). Sixth, GenAI-based screening results can be used for parallel independent reliability

[18] Here are but a few of the famous catchy titles: "To Err is Human: Building a Safer Health System" (Donaldson et al., 2000), "Guns, Germs, and Steel: The Fates of Human Societies" (Diamond, 1999), or "The Black Swan: The Impact of the Highly Improbable" (Taleb, 2007).

assessment, especially in single-authored review papers (Templier and Paré 2018), or for prioritizing screening activities (Syriani et al. 2024; van de Schoot et al. 2021). Finally, text summarization may even allow researchers to modify screening criteria and understand whether and how conclusions would change. If effective strategies of using GenAI for this purpose can be developed, this would allow researchers to complete qualitative robustness checks and validate some of the more challenging and consequential methodological choices. In addition, such work could show how screening criteria emerge from a mutually informative process iterating between humans and GenAI-based machines, as well as search, screen, and reading activities (Boell and Cecez-Kecmanovic 2014).

Table 7. Prompt for language translation in the screening process

GenAI-Capability	Text translation
Prompting Strategy	Zero-shot prompting
Requirements	Use GROBID to convert PDF documents to TEI format and provide the TEI (xml) files as an input to the LLM [7, 19]
Prompt Example	Read each xml document, which has the namespace http://www.tei-c.org/ns/1.0. Extract the following items: - title, which is in TEI/teiHeader/fileDesc/titelStmt/title (display in title case) - abstract, which is in TEI/teiHeader/profileDesc/abstract/div (using all p tags) - keywords, which are in TEI/teiHeader/profileDesc/textClass/keywords Translate the abstract to English (if necessary). Arrange all results in a Markdown table. Add a "screening" column.
Initial Evaluation[8]	

[19] Please note that the URL (*http://www.tei-c.org/ns/1.0*) is a namespace identifier and is not associated with a document that can be accessed through the browser.

Outcome	Successful translation of 5 out of 5 abstracts (GPT-4o), 5 out of 5 abstracts (Claude 3.5 Sonnet), and 5 out of 5 abstracts (Gemini 1.5 Pro)

Quality Assessment

Formal quality assessment is particularly relevant for theory-testing reviews, including meta-analyses and qualitative systematic reviews (Templier and Paré 2018). GenAI may assist with basic evaluations related to the methodological aspects of research studies, including the analysis of study designs, sample sizes, data collection methods, and statistical techniques (see Table 8, for an example). Future work may also show whether these models can be used effectively to identify potential flaws, biases, or limitations in the selected studies and flag them for further review by human researchers. Additionally, GenAI can help identify questionable research practices or logical errors (Habernal et al. 2018). It may also be leveraged to augment manual human quality assessments by ensuring consistency across multiple reviewers. By analyzing the assessments provided by different reviewers, the AI model can identify discrepancies or inconsistencies in the evaluation criteria or ratings, allowing for resolution and alignment.

Table 8. Prompt for the basic evaluation of the methodological approach

GenAI-Capability	**Text analysis and recommendations**
Prompting Strategy	Zero-shot prompting
Requirements	LLMs with file upload and large context window (> 100,000 tokens)
Prompt Example	**Upload a paper (PDF)** [7] Your task is to analyze the provided research study and identify its study design (e.g., experiment, case study, survey, archival study), sample size, data collection methods, and statistical analyses. Present these four characteristics in a Markdown table.
Initial Evaluation[8]	
Outcome	Convincing evaluation of 2 out of 5 methodological approaches (GPT-4o), 1 out of 5 methodological approaches (Claude 3.5 Sonnet), and 1 out of 5 methodological approaches (Gemini 1.5 Pro)

In the future, one potential application of GenAI may be conducting parallel independent assessments of study quality. The methodological literature recommends multiple independent

assessors evaluate the quality of studies included in a review (Templier and Paré 2018). GenAI models, with few-shot prompting, can perform these independent assessments, providing an additional layer of evaluation alongside human reviewers (Weber 2024). Furthermore, GenAI models could play a role in identifying, and refining the criteria used for quality assessment. By analyzing large datasets of literature reviews and their associated quality assessments, this involves identifying patterns or best practices in the assessment criteria and processes.

Data Extraction

In the data extraction activities, it is particularly instructive to consider the *jagged frontier* of GenAI. This refers to the observation that, instead of leading to consistent improvements across tasks, GenAI can have unpredictable effects when "tasks that appear to be of similar difficulty may either be performed better or worse by humans using AI" (Dell'Acqua et al. 2023, p. 8). For example, while GenAI may perform reliably when extracting explicit characteristics, such as sample sizes or participant demographics, yet fail in closely related situations requiring greater interpretive judgment, such as inferring an author's implicit epistemological stance or interpreting complex robustness checks in the reported results.

For descriptive reviews, where the goal is to summarize research findings, GenAI has already demonstrated its effectiveness in creating concise and accurate summaries. Importantly, these AI-generated summaries can be tailored to the researcher's specific needs, focusing on particular themes or methodological characteristics, or adjusted to suit different target audiences (see Table 9, for an example).

Table 9. Prompt for chain-of-density summarization

GenAI-Capability	Text summarization
Prompting Strategy	Chain-of-thought prompting
Requirements	LLMs with file upload and large context window (> 100,000 tokens)
Prompt Example	**Upload a paper (PDF)** [7] You will generate increasingly concise, entity-dense summaries of the above article. The summaries should be written for an academic audience. Repeat the following 2 steps 5 times. - Step 1. Identify 1-3 informative entities (";" delimited) from the article which are missing from the previously generated summary. - Step 2. Write a new, denser summary of identical length which covers every entity and detail from the previous summary plus the missing entities. A missing entity is: - Relevant: to the main story. - Specific: descriptive yet concise (5 words or fewer). - Novel: not in the previous summary. - Faithful: present in the article. Anywhere: located anywhere in the article.
Initial Evaluation[8]	
Outcome	Best performing prompt (after four chain of density steps) that maximize entity density and surpass human summaries (Adams et al. 2023)

While GenAI may be less useful for data extraction in review aimed at theory-development, it shows great promise for theory-testing reviews, such as meta-analyses, which rely on structured data extraction. GenAI can extract structured data from text, including study characteristics (Dagdelen et al. 2024), and potentially correlations and effect sizes from primary studies. These GenAI classification capabilities may be most useful for small or emergent subject areas where no validated ML models for classification exist. Before GenAI models, developing a text classifier for a non-standard problem, including the annotation of a training dataset, required a substantial amount of work. With GenAI, getting up and running with a classifier may become much easier (see Table 10, for an example).

Table 10. Python pseudocode for structured data extraction from tables

GenAI-Capability	Data extraction
Prompting Strategy	Zero-shot prompting
Requirements	LLMs with file upload and large context window (> 100,000 tokens)
Prompt Example	**Upload a paper (PDF)** [7] 1. Define utility functions: • md_to_df(markdown_text): Converts markdown table text to a pandas DataFrame. • extract_table_from_image(url): Extracts table data from an image at the given URL and returns as markdown text. 2. Define the MarkdownDataFrame data structure: • Use pandas.DataFrame as the base structure. • Apply a BeforeValidator that converts markdown text to a DataFrame (md_to_df function). • Apply a PlainSerializer to convert a DataFrame to markdown text (using DataFrame.to_markdown() method). • Define JSON schema for validation. 3. Define the Table class with two attributes: caption and dataframe: • caption: String to store the table's caption. • dataframe: Stores the table data as a MarkdownDataFrame, which is essentially a pandas DataFrame that can serialize to/from markdown. 4. Main process to extract and represent a table from an image: • Call extract_table_from_image(url) to extract the markdown representation of the table from the image. • Create an instance of the Table class, setting caption as needed and dataframe as the markdown representation converted to a DataFrame. • Use the Table instance to manipulate or access the table's data and caption. • To serialize the Table instance's dataframe back to markdown, use the PlainSerializer functionality implicitly via the class's structure.
Initial Evaluation[8]	
Outcome	Successful data extraction from 4 out of 5 tables (GPT-4o), 5 out of 5 tables (Claude 3.5 Sonnet), and 4 out of 5 tables (Gemini 1.5 Pro)

As GenAI continues to improve, with expanding context windows and enhanced ability to generate structured and reproducible output (Dagdelen et al. 2024), we can envision GenAI automating even more complex data extraction tasks. For instance, it may become possible for

GenAI to extract correlation tables, compile effect sizes from a sample of primary studies, and assist in automatically conducting meta-analyses. This could significantly reduce the time and effort required for theory-testing reviews (Li et al., 2026), allowing researchers to focus on other steps of the process or even entirely different types of reviews with more substantial interpretation and synthesis requirements.

Data Analysis

In data analysis and code development tasks, GenAI has demonstrated remarkable capabilities (Peng et al. 2023) and tools like MAXQDA have started to integrate LLM capabilities. Generally, one of the most popular use cases of GenAI—text generation—may be leveraged across all types of reviews, for example to draft descriptive summaries based on research papers in narrative, descriptive or scoping reviews as well as to assist in editing and proofreading tasks for theoretical reviews (Huang and Tan 2023; Skarlinski et al. 2024). Although some have argued that purely text-generative tools may not be capable of supporting evidence-aggregating studies (Rahman et al. 2023; Schryen et al. 2024), literature reviews that involve structured analyses such as meta-analyses may benefit from GenAI by developing code for analyses or data visualizations (see Table 11, for an example). Perhaps most significantly, recent work leverages GenAI to support dynamic, real-time theory testing reviews that automatically update as new studies are published, ensuring researchers always have a synthesis of the latest evidence available (Li et al., 2026). In this manner, GenAI can unleash a new form of publishing whereby a paper is a living document, updated in-vivo (on the publishing platform) as new evidence emerges.

Table 11. Prompt to develop Python code for a meta-analysis

GenAI-Capability	Code generation
Prompting Strategy	Zero-shot prompting
Requirements	LLMs
Prompt Example	As a Python programming and statistical analysis expert with a detailed understanding of conducting meta-analysis in Python, you are tasked with generating Python code that aligns with the following steps: - Step 1: Install the PythonMeta (V.1.26) package and read a dataset. The dataset is sitting in the same file directory as the Python scripts. - Step 2: Generate main results by selecting binary outcome and Risk Ratio as the desired effect size. Run both fixed-effect and random-effects models, choosing MH for fixed-effect and DL for the random-effects models. Generate forest plots and funnel plots. - Step 3: Assess the impact of missing data. After cleaning the dataset, label the studies with missing and non-missing patients and analyze them as subgroups. Implement missing data imputation methods including Available Case Study (ACS), Imputed Case Analysis (ICA), and best and worst-case scenarios. Run a separate random-effects model with IV method on each and generate relevant forest plots. - Step 4: Evaluate the small study effect, assess the asymmetry of the funnel plots, and perform Egger's test using Statsmodels linear regression. Remember to format the responses in a clear and precise format. Output tables when possible. Keep your tone professional and instructional, ensuring the generated Python code adheres to best practices for readability and efficiency.
Initial Evaluation[8]	
Outcome	Generation of working meta-analysis code with minor fixes (GPT-4o), working meta-analysis code on first attempt (Claude 3.5 Sonnet), and working meta-analysis code with minor fixes (Gemini 1.5 Pro)

On the other end of the spectrum, for theory-building reviews with less strict data analysis schemas and inductive analyses, the support from AI may be limited to writing assistance, as these tasks require a higher level of conceptual understanding and idea generation. Nonetheless, the interactional quality of chatbot implementations of GenAI can support idea generation and refinement through, for example, Socratic-style argumentation about emerging theoretical ideas which may satisfy criteria for "the epistemic community values of argumentation" (Ngwenyama

and Rowe 2024, p. 123), as illustrated in the example prompt below (see Table 12). Looking ahead, GenAI will play a more substantive role especially in inductive, theory-building work by acting as a co-researcher that summarizes the researcher's memos identifying gaps in understanding, identifying key concepts in research papers and mapping conceptual relationships. It could thereby complement existing approaches primarily applicable to hypothetico-deductive research (Li et al. 2020).

Table 12. Prompt to reframe theoretical questions based on Socratic argumentation

GenAI-Capability	Dialogue and conversation
Prompting Strategy	Exploratory prompting
Requirements	LLMs with file upload and large context window (> 100,000 tokens)
Prompt Example	**Upload a selection of relevant papers (PDFs)** [7] You are an AI assistant capable of having in-depth Socratic style conversations on a wide range of topics. Your goal is to ask probing questions to help the user critically examine their beliefs and perspectives on the attached paper. Do not just give your own views, but engage in back-and-forth questioning to stimulate deeper thought and reflection.
Initial Evaluation[8]	
Outcome	Adapted from best performing prompt (without extra knowledge) that involves structured conversation, encompassing review, heuristic, rectification, and summarization (Ding et al. 2024)

Reflections, Opportunities, Challenges, and Open Questions

The capabilities of GenAI to assist with literature reviews are already impressive and continue to improve, as companies and even countries begin to compete to create better foundational and specialized GenAI models. At the same time, important questions and opportunities related to methodological and technological challenges and the future of scientific progress must be raised, to ensure the use of GenAI tools for literature reviews is effective, but also responsible. In this section we aim to present a balanced outlook by highlighting the positive potential of GenAI for literature reviews while also critically questioning some developments that could undermine

long-term scientific innovation and creativity or pose risks from opening up scientific processes too broadly. In the following, we summarize our own findings, thereby providing a backdrop for more long-term reflections, including the potential impact of GenAI on scientific progress, types of review and the technological challenges of GenAI that should be overcome to unlock even greater potential of this transformative technology.

Discussion of our Findings

Despite the caveats and limitations, our analysis reveals considerable promises of GenAI, which is highly capable of augmenting and even, in some cases, entirely automating activities of the literature review process. Both generic (e.g., ChatGPT, Claude, Gemini) and specialized (e.g., Consensus, Elicit) tools should be actively considered by researchers on most review projects. At the same time, the tools are best thought of as methodological co-pilots, rather than wholesale replacements of manual human effort.

Some activities of the literature review process appear to be especially amenable to support or in some cases, full automation, with GenAI. Thus, problem formulation can be greatly enhanced by GenAI as it excels at navigating ambiguity, providing a sense of present-day developments and improving conceptual understanding of early-stage literature exploration, at an unprecedented scale. Similarly, for literature searches, GenAI is particularly apt at supporting or supplementing exploratory activities.

For some activities, GenAI can be incredibly helpful, but often requires careful and precise prompts, and does not always outperform manual or traditional AI-driven initiatives. If the aim is to maintain maximal transparency and controls, manual effort or carefully designed and meticulously validated (Ethayarajh and Jurafsky 2020; Larsen et al. 2025) custom AI classification models should be preferred (Wagner et al. 2022). Here we observe a perennial

tension between rigor and scale, known in other settings, as for example the trade-off between accuracy and completeness, precision and recall, or internal and external validity.

The generic abilities of GenAI to summarize content at scale, and transform content presentation (e.g., creating tables, graphs), enhance researchers' ability to understand and communicate the findings. The tools also permit the evolution and enhancement of the methods underlying literature reviews. For example, GenAI may permit qualitative robustness checks and hence validate some of the methodological choices, which are rarely validated at the moment (e.g., the search strategy or screening criteria). Similarly, GenAI may be used to conduct parallel independent data extraction, quality control, screening and search activities which can be compared to manual efforts. Incorporating these possibilities into literature review methods is an exciting frontier for research.

The present-day capabilities and even greater future potential of GenAI have profound implications for how GenAI may shape the trajectory of scientific progress in both beneficial and detrimental ways that need to be carefully managed. Next, we consider the broader effect on long-term scientific innovation and progress.

GenAI and Literature Review Types

Although GenAI shows considerable potential, its utility varies significantly depending on the type of literature review and the specific demands of each review stage. To explore this, we present four scenarios—*Obsolescence*, *Varying Degrees of Augmentation*, *Inadequacy*, and *New Trajectories*—each representing a unique pathway through which GenAI could transform literature review practices. These scenarios highlight GenAI's impact on different types of reviews, the changes it may introduce to specific review steps, and the broader implications for literature review methodologies.

In the *Obsolescence* scenario, GenAI advancements lead to the partial or complete replacement of certain types of literature reviews. Specifically, GenAI's summarization, automated synthesis, and bibliometric capabilities make some reviews—such as descriptive reviews and bibliometric studies—obsolete, particularly those focused on categorizing and summarizing existing literature. GenAI's ability to rapidly collect, classify, and synthesize large datasets reduces the need for manual effort in these review types. The activities most affected by this scenario include search and screening, where GenAI may fully automate or significantly streamline these processes, and synthesis, where AI can categorize and summarize literature with minimal human intervention.

The *Varying Degrees of Augmentation* scenario captures the spectrum of GenAI's potential impact on literature reviews, where GenAI serves as a supportive tool rather than a replacement. In this scenario, GenAI capabilities are selectively applied based on the complexity and demands of each review stage. For instance, narrative reviews, scoping reviews, and meta-analyses may benefit from GenAI in tasks like exploratory searches, screening, and data extraction, while human oversight remains essential for data synthesis and interpretation to maintain rigor and accuracy. GenAI's role in this scenario is to enhance traditional review processes by increasing efficiency and reducing researchers' time demands, allowing them to focus on high-level analysis and interpretation. This collaborative model underscores the need for researchers to retain methodological expertise while leveraging AI effectively.

In the *Inadequacy* scenario, GenAI tools, despite their capabilities, remain insufficient for certain types of literature reviews. Reviews requiring substantial theoretical, conceptual, or interpretive synthesis—such as theoretical reviews, conceptual reviews, meta-narrative reviews, meta-ethnographies, and critical reviews—demand deep contextual understanding and interpretative skills that GenAI currently lacks. Here, GenAI might minimally assist in exploratory search and preliminary reading, but it falls short in synthesis and critical interpretation. These review types

depend on researchers' interpretative lenses and subjective insights, which cannot be replicated by generative models alone. This scenario underscores GenAI's limitations, particularly when interpretative depth and nuanced analysis are required, and reinforces the ongoing importance of human expertise in qualitative and theoretical reviews. It highlights the value of "human-in-the-loop" models, where AI aids in preliminary tasks but requires significant human oversight for rigorous interpretation.

Finally, the *New Trajectories* scenario envisions GenAI as a catalyst for innovation, enabling novel types of literature reviews or expanding the scope of traditional reviews in ways previously infeasible. This scenario is especially relevant for interdisciplinary reviews that are simultaneously broad and in-depth, and those where researchers apply established theories to new, unrelated contexts (Gregor and Hevner, 2013). GenAI's potential for translational capabilities—its ability to bridge disciplinary knowledge and generate cross-disciplinary insights—allows researchers to engage with literature beyond their immediate field, fostering a new level of interdisciplinary integration. Moreover, GenAI could enable radically scaled review efforts, allowing researchers to analyze vast amounts of literature from multiple disciplines efficiently. By facilitating knowledge transfer across fields, this scenario could contribute to theoretical advancement and the development of interdisciplinary frameworks. However, it also requires researchers to remain vigilant in ensuring the accuracy and relevance of insights, particularly when working in unfamiliar domains.

In summary, these four scenarios illustrate the diverse ways in which GenAI could impact literature reviews. From the automation of descriptive reviews to selective augmentation in theory-testing reviews, limited applicability in interpretative reviews, and new opportunities in interdisciplinary synthesis, each scenario underscores a different facet of GenAI's transformative potential. As the researchers navigate these possibilities, they must balance GenAI's efficiencies

with critical oversight, ensuring methodological rigor and adapting to the evolving landscape of AI-supported research.

Drawing on established classifications of review types (Paré et al., 2015; Paré et al., 2023; Rowe, 2014; Schryen et al., 2020), Table 13 provides an at-a-glance overview of how each review type might leverage GenAI, noting that some types align better with specific scenarios than others. For instance, GenAI can significantly enhance meta-analyses by supporting systematic tasks such as data extraction, writing code for the meta-analytic regressions, and preliminary synthesis, aligning well with the *Varying Degrees of Augmentation* scenario. GenAI's ability to automate data handling processes, identify relevant studies, and summarize results increases the efficiency of meta-analyses, allowing researchers to focus on higher-level analysis and interpretation. However, for more complex synthesis and interpretation, particularly when assessing study heterogeneity or addressing nuanced methodological issues, human expertise remains essential. Therefore, while GenAI can streamline many aspects of meta-analyses, rigorous oversight and critical evaluation by researchers are still required to ensure accuracy and robustness in the final synthesis. As another illustration, GenAI can augment certain stages of critical reviews by helping with systematic aspects, such as locating and summarizing literature. However, the interpretive depth, evaluative focus, and critical perspective that define critical reviews place them primarily within the *Inadequacy* scenario, as these tasks require sophisticated human judgment (Block and Kuckertz, 2024). Thus, while GenAI may support some activities, the core critique remains a human-centered task.

Table 13. Alignment of Main Review Types with GenAI Integration Scenarios

Overarching goal	Review type	GenAI Integration Scenarios			
		Obsolescence	**Augmentation**	**Inadequacy**	**New trajectories**
Describing	Narrative	Unlikely	Moderate – supports search and summarization	Likely – requires interpretive synthesis	Unlikely
	Descriptive	Likely – potential for automation	Moderate – thematic extraction	Limited	Unlikely
	Scoping/mapping	Moderate – structured tasks	Likely – systematic search and categorization	Unlikely	Unlikely
Theory testing	Meta-analysis	Moderate – data extraction, synthesis	Likely – supports search, data extraction, and statistical synthesis	Moderate – requires oversight in complex synthesis	Moderate – potential for cross-disciplinary integration
	Systematic	Moderate – structured tasks	Likely – supports search, screening, and synthesis	Unlikely	Unlikely
	Umbrella	Moderate – high-level synthesis	Likely – search, screening, and summarization	Unlikely	Moderate – interdisciplinary synthesis
	Rapid	Moderate – high-level synthesis	Likely – prioritizes speed, search, screening	Unlikely	Unlikely
Theory building	Theoretical	Unlikely	Moderate – supports thematic organization	Likely – requires theoretical synthesis and interpretation	Moderate – potential for concept translation across disciplines
	Realist	Unlikely	Moderate – supports search, data extraction	Likely – requires context-sensitive interpretation	Moderate – future potential in contextual integration
Understanding	Meta-narrative	Unlikely	Likely – supports thematic organization	Likely – requires interpretative synthesis	Moderate – interdisciplinary knowledge translation

				across narratives	
	Critical	Unlikely	Moderate – supports initial stages	Likely – requires critical interpretive analysis	Unlikely
	Problematization	Unlikely	Moderate – supports initial stages	Likely – requires questioning assumptions	Unlikely

Addressing Technological Challenges of GenAI

While GenAI continues to impress with its capabilities, it also sometimes disappoints. In order to pave the way for even greater impact, several key limitations of modern GenAI need to be overcome. Considering our analysis of GenAI for literature reviews, we suggest fruitful opportunities for research that seeks to improve GenAI itself. We focus on two central issues: architectural and data-related challenges as areas of future research**.**

Architectural Challenges

Limitations of GenAI due to architectural issues of this technology pose significant hurdles for their effective application for literature reviews. One of the foremost challenges is the propensity of these models to produce hallucinations, i.e., generating information that is factually incorrect or not present in the source data. In the context of literature reviews, such hallucinations can lead to misrepresentation of research findings, citation of non-existent studies, or incorrect summarization of key concepts, thereby compromising the integrity of the review. To mitigate these issues, techniques like RAG have shown promising results (Li et al. 2024). RAG enhances factual accuracy by enabling models to access and reference external databases during generation. Another promising approach is the curation of knowledge graphs representing literature sources. In this approach, some of the especially critical semantics does not have to be extracted from literature sources and can be embedded directly. For example, statistical information reported (e.g., coefficients of structural equation models), or specific definitions (e.g.,

constructs, relationship among constructs), can be directly represented as knowledge graphs, ensuring precise incorporation of this information into GenAI representations. To support these developments, research is needed across a wide spectrum, ranging from the efficient collection and management of knowledge graphs (e.g., potentially supported by community-curated repositories, crowdsourcing, and other human in the loop approaches to ensure high accuracy of ground-based representations), along with continued work on graph-based knowledge embedding in large language models (Pan et al., 2024).

Understanding nuanced contexts and critical synthesis prevalent in academic literature are significant technological challenges that GenAI models struggle with. Conducting literature reviews not only requires a deep comprehension of domain-specific language (which can be provided with appropriate data and methods like RAG- as we noted before) and theoretical frameworks, but also requires critical thinking and reasoning. Studies in this domain have shown that GenAI models, while capable of analogical and moral reasoning, struggle with other reasoning tasks such as spatial reasoning (Agrawal 2023). General-purpose AI models might not capture important subtleties, leading to superficial synthesis or misinterpretations of critical concepts. This limitation hinders the ability of AI to fully assist in synthesizing complex scholarly work and may necessitate significant human oversight to correct and refine the outputs. As a remedy, specialized models trained on domain-specific corpora should be developed to address the issue of nuanced understanding. By tailoring models to specific fields researchers can improve the models' grasp of specialized terminology and complex concepts. Additionally, transfer learning and other approaches such as reinforcement learning with human feedback (RLHF) enable models to improve their abilities in critical thinking and reasoning. These approaches have resulted in more recently developed GenAI models such as OpenAI's O1 Preview, which is shown to substantially outperform humans in "systematic thinking, computational thinking, data literacy, creative thinking, scientific reasoning, and abstract reasoning." (Latif et al. 2024). Furthermore, recent research has found that the performance of the O1 model, which was developed utilizing advanced reinforcement learning techniques that significantly surpass traditional RLHF methods, consistently improves with increased reinforcement learning during

training (train-time compute) and with more time allocated for reasoning during inference (test-time compute) (Latif et al. 2024).

The lack of interpretability and transparency inherent in many AI models (such as deep learning – based LLMs) is a significant technological challenge. GenAI models often function as "black boxes," making it difficult for researchers to trace how specific outputs are generated from given inputs. This opaqueness is problematic in academic settings where the justification of conclusions and the reproducibility of results are essential. Researchers may find it challenging to trust the insights provided by AI models if they cannot understand the models' reasoning processes, which undermines the utility of these tools in conducting rigorous literature reviews. To address this issue, developments in explainable AI (XAI) are enhancing the interpretability of model outputs by offering insights into the decision-making processes of AI systems (Swamy et al. 2024). For instance, attention mechanisms and gradient-based attribution methods allow researchers to identify which parts of the input data the model focuses on when generating responses. This transparency helps researchers understand and trust the AI's contributions to literature reviews.

Handling multi-modal data introduces additional technological complexities. Multi-modal GenAI aims to process and integrate information from various sources such as text, images, graphs, and tables, which are commonly found in academic articles. However, effectively combining these different data modalities to generate coherent and meaningful analyses remains a significant challenge. The models may not accurately interpret visual data like charts or may fail to correlate information across modalities, resulting in incomplete or biased literature reviews. In the realm of multi-modal data processing, innovative architectures like Transformers with modality-specific encoders are improving the integration of diverse data types. Models such as OpenAI's CLIP (Contrastive Language-Image Pre-training) demonstrated promising performance in associating textual and visual information. These advancements can enhance the AI's ability to interpret and synthesize information from different formats commonly found in academic literature.

Computational resource demands also present a barrier. The sophisticated architectures of GenAI models require substantial processing power and memory. This requirement can limit accessibility for individual researchers or institutions with constrained resources, thereby impeding widespread adoption of these technologies in academia. The high costs associated with training and deploying such models can also divert funding from other critical research activities. Efforts to reduce computational requirements are also underway. Techniques like model pruning (Ma et al. 2023), quantization (Egashira et al. 2024), and knowledge distillation (Xu et al. 2024) help create smaller, more efficient models without significantly sacrificing performance. These approaches make it more feasible for researchers with limited resources to utilize advanced AI tools.

Data-related Challenges

As with any other data intensive artificial intelligent technology, the issues pertaining to data play an outsized role in the continued maturity of GenAI. There are many challenges and opportunities related to data management for GenAI in the context of literature reviews.

One of the significant issues is data access. A promise of GenAI for literature reviews is in its ability to expand the coverage of topics beyond what is humanly possible. However, the realization of this promise is being impeded by the inability of present tools to capture the entirety of the relevant literature. As a result, any literature review findings or analyses would be biased toward available sources. What is worse, some of the sources (e.g., ArXiv.org) while being relevant, may not guarantee a rigorous peer review process, and therefore may not be as reliable as the carefully curated sources in the inaccessible databases.

Design science researchers can address the many data-related issues from a multitude of perspectives. First, an opportunity exists to improve the GenAI's development routine to automatically ascertain the quality, bias and representativeness of the sources used (Parsons et al., 2025; Zhang et al., 2019), permitting the AI models to better leverage the training data in generating the responses. Effectively, this is the concept behind retrieval augmented generation (RAG), except the research focus here is on the upstream part of the AI training, such that the tools would become more sensitive to the varying levels of

data quality and representativeness. This issue, while a general one, is especially important for literature reviews, as relevant to a research question literature can drastically vary in its quality. Considering this variability, GenAI tools can offer different analyses, depending on the sensitivity of the research team to the sources and their quality levels (e.g., analysis on the entire available corpus, only the most reputable sources, gray literature, etc.).

Second, data management scholars can support the GenAI industry with solutions that can lessen the monetary burden of having to procure sources from paid databases. These approaches, for example, can draw upon research on differential privacy (Dwork, 2006) and information obfuscation (Liao et al., 2021), where only relevant information is made accessible and shared, to minimize such concerns related to copyright and intellectual property protection and reduce the information transfer volume.

Third, even scientific literature in highly curated, paid databases, is not necessarily bias free. Weber (2024), when considering AI as a tool for reviews, warns scientific disciplines exhibit often hard-to-detect entrenched biases. An important opportunity therefore is to develop systematic techniques to automatically identify these biases and make researchers aware of them. This work can leverage growing research on AI data bias identification and mitigation (Chen et al., 2024; Nazer et al., 2023; Tejani et al., 2024). Despite much progress, one overlooked opportunity is communicating bias to the user through a user interface, and ensuring that the user (e.g., scientist-in-training), knows how to appropriately account for the biases in the literature.

Finally, there is an important novel opportunity related to what we call *prompt data management*. Prompt data management is a new data management frontier that focuses on collection, curating, classification, and support for usage of effective prompts for GenAI. In our context, these prompts are tailored to literature reviews. Prompt curation requires its own considerations, different from the generic data management contexts. As we discussed and showed in our paper, in addition to curating the text of the prompt, certain properties and details of the prompt are important to capture and curate. Effective prompts for GenAI follow patterns which are not yet well-established and understood. For example, prompts for

literature review are often required to be issued in a particular sequence (as literature search is a complex and multi-phased process). Hence a challenge of prompt data management is understanding the patterns of effective prompts for literature reviews, classifying them appropriately so users can easily find those needed for their tasks, and developing accessible repositories for such prompts. To begin realizing this vision, we created a repository of literature review prompts, which will be updated continuously with recent prompts that are published and evaluated in scientific outlets. [20] Future research can study prompts data management to better understand and refine the practices for collecting and curating literature review prompts.

In conclusion, although some technological challenges may currently limit the full potential of GenAI for literature reviews, ongoing advancements and innovations are steadily overcoming these obstacles. As the technology matures, we can anticipate more reliable, interpretable, and accessible GenAI models that will significantly enhance the efficiency and depth of literature review.

General open questions

There are many open questions, beyond design of GenAI and the identification of effective prompts, including standards for human oversight, and reporting principles. More fundamentally, new answers are needed on how GenAI can enrich human understanding. Accordingly, researchers should explore possibilities of bringing GenAI to hermeneutic traditions and theory development reviews, fostering a new era of interdisciplinary research that bridges the gap between computational analysis and human interpretation. At the top of the human cognitive ability pyramid lie creativity, critical thinking, and complex problem solving. When paired with GenAI capabilities, these skills have the potential to enhance our understanding, interpretation, and synthesis of prior knowledge.

[20] https://fs-ise.github.io/gen-ai-lr-prompts/

Careful attention should also be given to the misuse of GenAI (cf. Susarla et al., 2023). In the context of literature reviews, the recent work of Tingelhoff et al. (2025) offers an instructive discussion on what may be considered legitimate use of GenAI for literature review, or as the authors put it, "*what we should allow GenAI to do"* (p.1). Akin to other research methods, the potential of GenAI misuse in submitted literature review papers raises challenging questions for editors and reviewers, who are confronted with rising submission numbers but lack effective means to detect the misuse of GenAI.

There is already a concerning "tendency to offload human cognition and intelligence to GenAI which can have potentially dysfunctional consequences that are, at this time, largely unknown" (Susarla et al. 2023, p. 405). We believe it is prudent to also seriously consider what GenAI means for broader scientific progress. While full consequences of using GenAI for literature reviews remain uncertain, quite likely, for some research teams, the level of innovation will spike following the use of GenAI for literature reviews, whereas for other teams, it may be to their detriment. This brings a research opportunity to understand when, and under what conditions, the negative or the positive tendency develops. Providing a comprehensive and contextualized answer to this question stands to benefit scientific progress and broadly human society, which depends on science and its development.

The utility of GenAI in propelling scientific inquiry, particularly in conducting literature reviews, significantly depends on the researchers' approach to integrating these technologies into their research efforts. Researchers endowed with a deep understanding of their investigative domains are aptly equipped to critically evaluate the outputs generated by tools such as ChatGPT and Gemini, aligning them with their domain knowledge to identify promising pathways to pursue in their research endeavors. Consequently, for seasoned scholars, GenAI holds the potential to substantially enhance research outcomes. Conversely, researchers with less prior exposure to the focal literature may encounter difficulties in accurately assessing the relevance and validity of GenAI-generated outputs, potentially leading to the exploration of less viable research avenues. In these instances, GenAI might inadvertently impede scientific progress even if the sheer volume of scientific papers grows. This may be another "the rich get richer, and the poor

get poorer" scheme. This interplay underscores the indispensable role of domain-specific knowledge in maximizing the benefits of GenAI, highlighting the imperative for a synergistic integration of researcher acumen and technology to foster scientific advancement. The good news is, the debates about ways to synergize humans and AI are now abound, e.g., in the context of future of human work, software development (Jain et al. 2021; Lukyanenko et al. 2025), the methodology of AI supported literature reviews can learn from and contribute to these debates.

Concluding Remarks

Current discussions on how GenAI could facilitate the conduct of literature reviews are characterized by the excitement of new opportunities, as well as cautious and critical commentaries. Building on the preceding commentary on AI-supported literature reviews (Wagner et al. 2022), we aim to develop a more substantive connection to the established methodological discourse, and offer a balanced view, suggesting for which tasks GenAI may be beneficial, and clarifying potential shortcomings. Currently, the design of research tools and services in this area is evolving rapidly, but effectively using GenAI to conduct review projects requires a particular set of skills, and a closer alignment with established methodological principles.

We hope this paper contributes to a constructive foundation for GenAI-supported literature reviews across science and in other settings (e.g., business, private), where making decisions based on prior literature is happening.

References

Adams, G., Fabbri, A. R., Ladhak, F., Lehman, E., & Elhadad, N. (2023). From sparse to dense: GPT-4 summarization with chain of density prompting. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. (4th New Frontier Summarization Workshop)*. (https://doi.org/ 10.18653/v1/2023.newsum-1.7)

Agrawal, S. (2023). "Are LLMs the master of all trades?: Exploring domain-agnostic reasoning skills of LLMs." *arXiv preprint arXiv:2303.12810*.

Aguinis, H., Ramani, R. S., and Alabduljader, N. (2023). "Best-practice recommendations for producers, evaluators, and users of methodological literature reviews," *Organizational Research Methods* 26(1), 46–76. (https://doi.org/10.1177/1094428120943281).

Alavi, M., Leidner, D. E., and Mousavi, R. (2024). "Knowledge management perspective of generative artificial intelligence," *Journal of the Association for Information Systems* 25(1), 1–12. (https://doi.org/10.17705/1jais.00859).

Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A. E. E., and Zayed, T. (2023). "Harnessing the power of ChatGPT for automating systematic review process: Methodology, case study, limitations, and future directions," *Systems* 11(7), 351. (https://doi.org/10.3390/SYSTEMS11070351).

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). "On the dangers of stochastic parrots: Can language models be too big? 🦜," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623. (https://doi.org/10.1145/3442188.3445922).

Berente, N., Seidel, S., and Safadi, H. (2019). "Research Commentary - Data-driven computationally intensive theory development," *Information Systems Research* 30(1), 50–64. (https://doi.org/10.1287/ISRE.2018.0774).

Biyela, S., Dihal, K., Gero, K. I., Ippolito, D., Menczer, F., Schäfer, M. S., and Yokoyama, H. M. (2024). "Generative AI and Science Communication in the Physical Sciences," *Nature Reviews Physics* 6(3), 162–165. (https://doi.org/10.1038/S42254-024-00691-7).

Block, J., & Kuckertz, A. (2024). What is the future of human-generated systematic literature reviews in an age of artificial intelligence? *Management Review Quarterly*, 1-6. (https://doi.org/ 10.1007/s11301-024-00471-8)

Boell, S. K., and Cecez-Kecmanovic, D. (2014). "A Hermeneutic Approach for Conducting Literature Reviews and Literature Searches," *Communications of the Association for Information Systems* (34), 257–286. (https://doi.org/10.17705/1CAIS.03412).

Boell, S. K., and Cecez-Kecmanovic, D. (2015). "On Being 'Systematic' in Literature Reviews in IS," *Journal of Information Technology* 30(2), 161–173. (https://doi.org/10.1057/JIT.2014.26).

Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., and Parrish, J. K. (2014). "Next Steps for Citizen Science," *Science* 343(6178), 1436–1437. (https://doi.org/10.1126/SCIENCE.1251554).

Booth, A. (2008). "Unpacking Your Literature Search Toolbox: On Search Styles and Tactics," *Health Information & Libraries Journal* 25(4), 313–317. (https://doi.org/10.1111/J.1471-1842.2008.00825.X).

Bornmann, L., Haunschild, R., and Mutz, R. (2021). "Growth Rates of Modern Science: A Latent Piecewise Growth Curve Approach to Model Publication Numbers from Established and New Literature Databases," *Humanities and Social Sciences Communications* 8(1), 1–15. (https://doi.org/10.1057/S41599-021-00903-W).

Bowman, S. R. (2023). *Eight Things to Know about Large Language Models*. (https://arxiv.org/abs/2304.00612).

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray,

Wagner G., Prester, J. Mousavi, R., Lukyanenko R., and Paré G. (2026). Generative Artificial intelligence for literature reviews, ***Journal of Information Technology***. *Open access*

S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). *Language Models Are Few-Shot Learners*. (https://arxiv.org/abs/2005.14165).

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S. (2023). "Sparks of Artificial General Intelligence: Early Experiments with Gpt-4," *arXiv Preprint arXiv:2303.12712*. (https://doi.org/10.48550/ARXIV.2303.12712).

Buchanan, J., Hill, S., and Shapoval, O. (2024). "ChatGPT Hallucinates Non-Existent Citations: Evidence from Economics," *The American Economist* 69(1), 80–87. (https://doi.org/10.1177/05694345231218454).

Burton-Jones, A., Boh, W. F., Oborn, E., and Padmanabhan, B. (2021). "Editor's Comments: Advancing Research Transparency at MIS Quarterly: A Pluralistic Approach," *MIS Quarterly* (45:2), pp. iii–xviii.

Castillo-Segura, P., Alario-Hoyos, C., Kloos, C. D., and Fernández Panadero, C. (2023). "Leveraging the Potential of Generative AI to Accelerate Systematic Literature Reviews: An Example in the Area of Educational Technology," in *2023 World Engineering Education Forum - Global Engineering Deans Council (WEEF-GEDC)*, pp. 1–8. (https://doi.org/10.1109/WEEF-GEDC59520.2023.10344098).

Chen, F., Wang, L., Hong, J., Jiang, J., & Zhou, L. (2024). Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. *Journal of the American Medical Informatics Association*, *31*(5), 1172-1183.

Chomsky, N., Roberts, I., and Watumull, J. (2023). "Noam Chomsky: The False Promise of ChatGPT," *The New York Times* (8).

Clark, K. (2020). "Electra: Pre-training text encoders as discriminators rather than generators." *arXiv preprint arXiv:2003.10555*.

Cooper, C. (2016). *Citizen Science: How Ordinary People Are Changing the Face of Discovery*, Abrams.

Daft, R. L. (1983). "Learning the Craft of Organizational Research," *Academy of Management Review* 8(4), 539–546. (https://doi.org/10.5465/AMR.1983.4284649).

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., and Jain, A. (2024). "Structured Information Extraction from Scientific Text with Large Language Models," *Nature Communications* 15(1), 1418. (https://doi.org/10.1038/S41467-024-45563-X).

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," *Harvard Business School Technology & Operations Mgt. Unit Working Paper* (24-013). (https://doi.org/10.2139/SSRN.4573321).

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171-4186). (https://doi.org/10.18653/v1/N19-1423)

Diamond, J. M. (1999). Guns, Germs, and Steel: The Fates of Human Societies.

Ding, Y., Hu, H., Zhou, J., Chen, Q., Jiang, B., & He, L. (2024). Boosting large language models with socratic method for conversational mathematics teaching. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management* (pp. 3730-3735). (https://doi.org/10.1145/3627673.3679881)

Dissanayake, I., Nerur, S. P., Lukyanenko, R., & Modaresnezhad, M. (Accepted, Jan 2025). The State-of-the-Art of Crowdsourcing Systems: A Computational Literature Review and Future Research Agenda Using a Text Analytics Approach. *Information & Management*, 62 (104098), pp. 1-14.

Donaldson, M. S., Corrigan, J. M., & Kohn, L. T. (Eds.). (2000). To err is human: building a safer health system. (https://doi.org/10.17226/9728)

Dwork, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming,* pp. 1-12.

Egashira, K., Vero, M., Staab, R., He, J., Vechev, M. (2024). "Exploiting LLM Quantization." *arXiv preprint arXiv:2405.18137*.

Wagner G., Prester, J. Mousavi, R., Lukyanenko R., and Paré G. (2026). Generative Artificial intelligence for literature reviews, ***Journal of Information Technology***. *Open access*

Ethayarajh, K., and Jurafsky, D. (2020). "Utility Is in the Eye of the User: A Critique of NLP Leaderboards," *ArXiv Preprint* ArXiv:2009.13888.

Eveleigh, A., Jennett, C., Lynn, S., and Cox, A. L. (2013). "I Want to Be a Captain! I Want to Be a Captain!: Gamification in the Old Weather Citizen Science Project," in *Proceedings of the First International Conference on Gameful Design, Research, and Applications*, pp. 79–82. (https://doi.org/10.1145/2583008.2583019).

Feuerriegel, S., Hartmann, J., Janiesch, C. and Zschech, P., (2024). Generative AI. *Business & Information Systems Engineering*, *66*(1), pp.111-126. (https://doi.org/10.1007/s12599-023-00834-7)

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., and Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, *25*(3), 277-304. (https://doi.org/10.1080/15228053.2023.2233814)

Gemini Team Google: Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., and others (2023). Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805. (https://doi.org/10.48550/arXiv.2312.11805)

George, A. S., and George, A. H. (2023). "A Review of ChatGPT AI's Impact on Several Business Sectors," *Partners Universal International Innovation Journal* 1(1), 9–23.

Gigerenzer, G. E., Hertwig, R. E., and Pachur, T. E. (2011). *Heuristics: The Foundations of Adaptive Behavior*, Oxford, UK: Oxford University Press.

Gilovich, T., Griffin, D., and Kahneman, D. (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge, UK: Cambridge University Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems* 27, 2672-2680.

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., and Brauner, J. (2024). "Thousands of AI Authors on the Future of AI," *arXiv Preprint arXiv:2401.02843*.

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly* 37(2), 337-355. (https://doi.org/10.25300/MISQ/2013/37.2.01)

Gusenbauer, M., and Haddaway, N. R. (2021). "What Every Researcher Should Know about Searching–Clarified Concepts, Search Advice, and an Agenda to Improve Finding in Academia," *Research Synthesis Methods* 12(2), 136–147. (https://doi.org/10.1002/JRSM.1457).

Habernal, I., Pauli, P., and Gurevych, I. (2018). "Adapting Serious Game for Fallacious Argumentation to German: Pitfalls, Insights, and Best Practices," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Yao, Y., Zhang, A., Zhang, L., and others. (2021). "Pre-Trained Models: Past, Present and Future," *AI Open* (2), 225–250. (https://doi.org/10.1016/J.AIOPEN.2021.08.002).

Hendricks, F. (2024). "Business Value of Generative AI Use Cases," *Journal of AI, Robotics & Workplace Automation* 3(1), 47–54.

Hevner, A., Parsons, J., Brendel, A. B., Lukyanenko, R., Tiefenbeck, V., Tremblay, M. C., and vom Brocke, J. (2024). "Transparency in Design Science Research," *Decision Support Systems* (182:1), pp. 1–19. (https://doi.org/10.1016/j.dss.2024.114236)

Higgins, J., Thomas, J., Chandler, J., Cumpston, M., Li, T., and Page, V., MJ and Welch. (2023). *Cochrane Handbook for Systematic Reviews of Interventions*. (www.training.cochrane.org/handbook).

Ho, J., Jain, A. N., and Abeel, P. (2020). Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 6840–6851.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., et al. (2022). Training compute-optimal large language models. arXiv preprint arXiv:2203.15556. (https://doi.org/10.48550/arXiv.2203.15556)

Huang, J., and Tan, M. (2023). "The Role of ChatGPT in Scientific Communication: Writing Better Scientific Review Articles," *American Journal of Cancer Research* 13(4).

Wagner G., Prester, J. Mousavi, R., Lukyanenko R., and Paré G. (2026). Generative Artificial intelligence for literature reviews, ***Journal of Information Technology***. *Open access*

Hubert, K. F., Awa, K. N., and Zabelina, D. L. (2024). "The Current State of Artificial Intelligence Generative Language Models Is More Creative Than Humans on Divergent Thinking Tasks," *Scientific Reports* 14(1), 3440. (https://doi.org/10.1038/S41598-024-53303-W).

Hutson, M. (2023). "Hypotheses Devised by AI Could Find 'Blind Spots' in Research," *Nature*. (https://doi.org/10.1038/D41586-023-03596-0).

Hwang, K., and Sung, W. (2015). Single stream parallelization of generalized LSTM-like RNNs on a GPU. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 1047-1051). (https://doi.org/10.1109/ICASSP.2015.7178129)

Irvin, R. A., and Stansbury, J. (2004). "Citizen Participation in Decision Making: Is It Worth the Effort?" *Public Administration Review* 64(1), 55–65. (https://doi.org/10.1111/J.1540-6210.2004.00346.X).

Jain, H., Padmanabhan, B., Pavlou, P. A., & Raghu, T. S. (2021). Editorial for the special section on humans, algorithms, and augmented intelligence: The future of work, organizations, and society. *Information Systems Research*, *32*(3), 675-687.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys* 55(12), 1–38. (https://doi.org/10.1145/3571730).

Jin, Di, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. (2021). "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences* 11(14), 6421.

Johns, G. (2001). "In Praise of Context," *Journal of Organizational Behavior* 22(1), 31–42.

Johns, G. (2017). "Reflections on the 2016 Decade Award: Incorporating Context in Organizational Research," *Academy of Management Review* 42(4), 577–595. (https://doi.org/10.5465/AMR.2017.0044).

Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., and McHardy, R. (2023). Challenges and applications of large language models. arXiv preprint arXiv:2307.10169. (https://doi.org/10.48550/arXiv.2307.10169)

Khlaif, Z. N., Mousa, A., Hattab, M. K., Itmazi, J., Hassan, A. A., Sanmugam, M., and Ayyoub, A. (2023). The potential and concerns of using AI in scientific research: ChatGPT performance evaluation. *JMIR Medical Education*, *9*, e47049. (https://doi.org/10.2196/47049)

Kuhn, T. S. (1997). *The Structure of Scientific Revolutions*, University of Chicago press Chicago.

Kuroiwa, T., Sarcon, A., Ibara, T., Yamada, E., Yamamoto, A., Tsukamoto, K., and Fujita, K. (2023). "The Potential of ChatGPT as a Self-Diagnostic Tool in Common Orthopedic Diseases: Exploratory Study," *Journal of Medical Internet Research* (25), e47621. (https://doi.org/10.2196/47621).

Larsen, K. R., and Bong, C. H. (2016). "A Tool for Addressing Construct Identity in Literature Reviews and Meta-Analyses," *MIS Quarterly* 40(3). (https://doi.org/10.25300/MISQ/2016/40.3.01).

Larsen, K. R., Hovorka, D., Dennis, A. R., and West, J. D. (2019). "Understanding the Elephant - the Discourse Approach to Boundary Identification and Corpus Construction for Theory Review Articles," *Journal of the Association for Information Systems* 20(7), 887–927. (https://doi.org/10.17705/1JAIS.00556).

Larsen, K. R., Lukyanenko, R., Mueller, R., Storey, V. C., Parsons, J., Vander Meer, D., and Hovorka, D. (2025). "Validity in Design Science," *MIS Quarterly*, pp. 1–40.

Latif, E., et al. (2024). "A systematic assessment of openai o1-preview for higher order thinking in education," *arXiv preprint arXiv:2410.21287*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., and others. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems* (33), 9459–9474. (https://doi.org/10.5555/3495724.3496517).

Li, L., Mathrani, A., Susnjak, T. (2026). Transforming evidence synthesis: A systematic review of the evolution of automated meta-analysis in the age of AI. *Research Synthesis Methods*. Published online:1-48. doi:10.1017/rsm.2025.10065

Wagner G., Prester, J. Mousavi, R., Lukyanenko R., and Paré G. (2026). Generative Artificial intelligence for literature reviews, ***Journal of Information Technology***. *Open access*

Li, J., Larsen, K. R., and Abbasi, A. (2020). "TheoryOn - a Design Framework and System for Unlocking Behavioral Knowledge Through Ontology Learning," *MIS Quarterly* (44:4). (https://doi.org/10.25300/MISQ/2020/15323).

Li, Jiarui, Ye Yuan, and Zehua Zhang. (2024) "Enhancing LLM factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases," *arXiv preprint arXiv:2403.10446*.

Liao, P., Zhao, H., Xu, K., Jaakkola, T., Gordon, G. J., Jegelka, S., & Salakhutdinov, R. (2021). Information obfuscation of graph neural networks. In *International conference on machine learning*, pp. 6600-6610.

Lindberg, A. (2020). "Developing Theory Through Integrating Human and Machine Pattern Recognition," *Journal of the Association for Information Systems* 21(1), 7. (https://doi.org/10.17705/1JAIS.00593).

Lukyanenko, R., Parsons, J., Wiersma, Y., and Maddah, M. (2019). "Expecting the Unexpected: Effects of Data Collection Design Choices on the Quality of Crowdsourced User-Generated Content," *MIS Quarterly* 43(2), 634–647. (https://doi.org/10.25300/MISQ/2019/14439).

Lukyanenko R., Samuel B. M., Tegarden D., Larsen K. R., Jabbari A., Abd El Aziz, R., Almeida J. P. A., Amrollahi, A., Beard, J.W., Bellatreche L., Bork, D., vom Brocke, J., Cabot, J., Castellanos, A., Gerber, A., Green, P., Grover, A., Guizzardi, G., Hertelendy, A., Kahlon, Y., Karlapalem, K., Khatri, V., Maass, W., Maurer, C., Mueller, R. M., Mylopoulos J., Nishant, R., Ogunseye, S., Paré, G., Parsons J., Pastor, O., Prester, J., Proper, H.A., Ralyte, J., Sadiq, S., Schuff, D., Siau, K., Snoeck M., Storey, V.C., Tremblay, M.C., Trujillo J., van der Meer, D., Venkataraman, R., Wagner, G., Wiedemann, A., Woo, C., Yu, E., Yue, W. T., Zhao, L., (2025). Manifesto for Software Development and Modeling in the Age of Artificial Intelligence. Available at SSRN: https://ssrn.com/abstract=5881104 or http://dx.doi.org/10.2139/ssrn.5881104

Ma, Xinyin, Gongfan Fang, and Xinchao Wang. (2023). "LLM-pruner: On the structural pruning of large language models," *Advances in Neural Information Processing Systems* 36: 21702-21720.

McGowan, A., Gui, Y., Dobbs, M., Shuster, S., Cotter, M., Selloni, A., Goodman, M., Srivastava, A., Cecchi, G. A., and Corcoran, C. M. (2023). "ChatGPT and Bard Exhibit Spontaneous Citation Fabrication During Psychiatry Literature Search," *Psychiatry Research* (326), 115334. (https://doi.org/10.1016/J.PSYCHRES.2023.115334).

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., and Roth, D. (2023). "Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey," *ACM Computing Surveys* 56(2), 1–40. (https://doi.org/10.1145/3605943).

Müller-Bloch, C., and Kranz, J. (2015). "A Framework for Rigorously Identifying Research Gaps in Qualitative Literature Reviews," in *Proceedings of the International Conference on Information Systems*. (http://aisel.aisnet.org/icis2015/proceedings/ResearchMethods/2).

Nazer, L. H., Zatarah, R., Waldrip, S., Ke, J. X. C., Moukheiber, M., Khanna, A. K., ... & Mathur, P. (2023). Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health*, *2*(6), e0000278.

Nelson, L. K. (2020). "Computational Grounded Theory: A Methodological Framework," *Sociological Methods & Research* 49(1), 3–42. (https://doi.org/10.1177/0049124117729703).

Ngwenyama, O., and Rowe, F. (2024). "Should We Collaborate with AI to Conduct Literature Reviews? Changing Epistemic Values in a Flattening World," *Journal of the Association for Information Systems* 25(1), 122–136. (https://doi.org/10.17705/1JAIS.00869).

Northcraft, G. B., and Neale, M. A. (1987). "Experts, Amateurs, and Real Estate: An Anchoring-and-Adjustment Perspective on Property Pricing Decisions," *Organizational Behavior and Human Decision Processes* 39(1), 84–97. (https://doi.org/10.1016/0749-5978(87)90046-X).

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, 372. (https://doi.org/10.1136/bmj.n71)

Wagner G., Prester, J. Mousavi, R., Lukyanenko R., and Paré G. (2026). Generative Artificial intelligence for literature reviews, ***Journal of Information Technology***. *Open access*

Palani, S., Naik, A., Downey, D., Zhang, A. X., Bragg, J., & Chang, J. C. (2023). Relatedly: Scaffolding literature reviews with existing related work sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-20).

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Paré, G., Tate, M., Johnstone, D., and Kitsiou, S. (2016). "Contextualizing the Twin Concepts of Systematicity and Transparency in Information Systems Literature Reviews," *European Journal of Information Systems* 25(6), 493–508. (https://doi.org/10.1057/S41303-016-0020-3).

Paré, G., Trudel, M.-C., Jaana, M., and Kitsiou, S. (2015). "Synthesizing Information Systems Knowledge - a Typology of Literature Reviews," *Information & Management* 52(2), 183–199. (https://doi.org/10.1016/J.IM.2014.08.008).

Paré, G., Wagner, G. & Prester, J. (2024). "How to Develop and Frame Impactful Review Articles: Key Recommendations," *Journal of Decision Systems* 33(4), pp. 566-582. (https://doi.org/10.1080/12460125.2023.2197701)

Parsons, J., Lukyanenko, R., Greenwood, B., & Cooper, C. (2025). Understanding and Improving Data Repurposing. *MIS Quarterly*, pp.1-27. (https://doi.org/10.25300/MISQ/2025/18361)

Peng, S., Kalliamvakou, E., Cihon, P., and Demirer, M. (2023). *The Impact of AI on Developer Productivity: Evidence from GitHub Copilot*. (https://arxiv.org/abs/2302.06590).

Poisson, A. C., McCullough, I. M., Cheruvelil, K. S., Elliott, K. C., Latimore, J. A., and Soranno, P. A. (2020). "Quantifying the Contribution of Citizen Science to Broad-Scale Ecological Databases," *Frontiers in Ecology and the Environment* 18(1), 19–26. (https://doi.org/10.1002/FEE.2128).

Prester, J., Wagner, G., Schryen, G., & Hassan, N. R. (2021). Classifying the ideational impact of information systems review articles: A content-enriched deep learning approach. *Decision Support Systems*, 140, 113432. (https://doi.org/10.1016/j.dss.2020.113432)

Qureshi, R., Shaughnessy, D., Gill, K. A. R., Robinson, K. A., Li, T., and Agai, E. (2023). "Are ChatGPT and Large Language Models 'the Answer' to Bringing Us Closer to Systematic Review Automation?" *Systematic Reviews* 12(1), 72. (https://doi.org/10.1186/S13643-023-02243-Z).

Radford, A., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer." *Journal of Machine Learning Research* 21(140), 1-67.

Rahman, M., Terano, H. J. R., Rahman, N., Salamzadeh, A., and Rahaman, S. (2023). "ChatGPT and Academic Research: A Review and Recommendations Based on Practical Examples," *Journal of Education, Management and Development Studies* 3(1), 1–12. (https://doi.org/10.52631/JEMDS.V3I1.175).

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., and Tomsett, R. (2022). "Deciding Fast and Slow: The Role of Cognitive Biases in AI-Assisted Decision-Making," *Proceedings of the ACM on Human-Computer Interaction*, pp. 1–22. (https://doi.org/10.1145/3512930).

Recker, J., Lukyanenko, R., Jabbari, M. A., Samuel, B. M., and Castellanos, A. (2021). "From Representation to Mediation: A New Agenda for Conceptual Modeling Research in A Digital World," *MIS Quarterly*. 45 (1), pp. 269-300.

"Rise of the Citizen Scientist." (2015). *Nature* (524:265). (https://doi.org/10.1038/524265A).

Rivard, S. (2024). "Unpacking the process of conceptual leaping in the conduct of literature reviews," *The Journal of Strategic Information Systems* 33(1), 1-8. (https://doi.org/10.1016/j.jsis.2024.101822)

Rivard, S., Constantiou, I., and Hsu, C. (2018). "Call for Proposals for Review Articles," *The Journal of Strategic Information Systems* 27(2), I–II. (https://doi.org/10.1016/S0963-8687(18)30157-4).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10684-10695.

Wagner G., Prester, J. Mousavi, R., Lukyanenko R., and Paré G. (2026). Generative Artificial intelligence for literature reviews, ***Journal of Information Technology***. *Open access*

Rowe, F. (2014). What literature review is not: diversity, boundaries and recommendations. *European Journal of Information Systems*, 23(3), 241-255. (https://doi.org/10.1057/ejis.2014.7)

Russo, C., Veronelli, L., Casati, C., Monti, A., Perucca, L., Ferraro, F., Corbo, M., Vallar, G., and Bolognini, N. (2021). "Explicit Motor Sequence Learning After Stroke: A Neuropsychological Study," *Experimental Brain Research* 239(7), 2303–2316. (https://doi.org/10.1007/S00221-021-06141-5).

Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E., and Akata, Z. (2023). "In-context impersonation reveals large language models' strengths and biases," in *Advances in Neural Information Processing Systems (NeurIPS 2023)*.

Sanderson, K. (2023). "AI Science Search Engines Are Exploding in Number—Are They Any Good?" *Nature* 616(7958), 639–640. (https://doi.org/10.1038/D41586-023-01273-W).

Schmeller, D. S., HENRY, P.-Y., Julliard, R., Gruber, B., Clobert, J., Dziock, F., Lengyel, S., Nowicki, P., Deri, E., Budrys, E. (2009). "Advantages of Volunteer-Based Biodiversity Monitoring in Europe," *Conservation Biology* 23(2), 307–316. (https://doi.org/10.1111/J.1523-1739.2008.01125.X).

Schryen, G., Marrone, M., and Yang, J. (2024). "Adopting Generative AI for Literature Reviews: An Epistemological Perspective," in *Proceedings of the 57th Hawaii International Conference on System Sciences*.

Schryen, G., Wagner, G., Benlian, A., and Paré, G. (2020). A knowledge development perspective on literature reviews: Validation of a new typology in the IS field. *Communications of the Association for Information Systems*, 46. (https://doi.org/10.17705/1CAIS.04607)

Skarlinski, M. D., Cox, S., Laurent, J. M., Braza, J. D., Hinks, M., Hammerling, M. J., Ponnapati, M., Rodriguez, S. G., & White, A. D. (2024). Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*.

Sohl-Dickstein, J., et al. (2015). Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *Proceedings of the 32nd International Conference on Machine Learning*, 2256–2265.

Storey, V. C., Zhao, J. L., Wei Thoo, Y., and Lukyanenko, R. (2025). Generative artificial intelligence: Evolving technology, growing societal impact, and opportunities for information systems research. *Information Systems Frontiers*, 1-22. (https://doi.org/10.1007/s10796-025-10581-7)

Sun, K., and Wang, R. (2025). "Systematic framework of application methods for large language models in language sciences," *arXiv preprint arXiv:2512.09552*.

Susarla, A., Gopal, R., Thatcher, J. B., and Sarker, S. (2023). "The Janus Effect of Generative AI: Charting the Path for Responsible Conduct of Scholarly Activities in Information Systems," *Information Systems Research* 34(2), 399–408. (https://doi.org/10.1287/ISRE.2023.ED.V34.N2).

Swamy, V., et al. (2024). "From Explanations to Action: A Zero-Shot, Theory-Driven LLM Framework for Student Performance Feedback," *arXiv preprint arXiv:2409.08027*.

Syriani, E., David, I., and Kumar, G. (2024). "Screening articles for systematic reviews with ChatGPT," *Journal of Computer Languages* 80, 101287. (https://doi.org/10.1016/j.cola.2024.101287).

Taleb, N. N. (2007). The black swan: The impact of the highly improbable. Random House.

Tang, X., Li, X., Ding, Y., Song, M., and Bu, Y. (2020). "The Pace of Artificial Intelligence Innovations: Speed, Talent, and Trial-and-Error," *Journal of Informetrics* 14(4), 101094. (https://doi.org/10.1016/J.JOI.2020.101094).

Tejani, A. S., Ng, Y. S., Xi, Y., & Rayan, J. C. (2024). Understanding and mitigating bias in imaging artificial intelligence. *RadioGraphics*, *44*(5), e230067.

Templier, M., and Paré, G. (2018). "Transparency in Literature Reviews - an Assessment of Reporting Practices Across Review Types and Genres in Top IS Journals," *European Journal of Information Systems* 27(5), 503–550. (https://doi.org/10.1080/0960085X.2017.1398880).

Temsah, O., Khan, S. A., Chaiah, Y., Senjab, A., Alhasan, K., Jamal, A., Aljamaan, F., Malki, K. H., Halwani, R., Al-Tawfiq, J. A. (2023). "Overview of Early ChatGPT's Presence in Medical Literature: Insights from a Hybrid Literature Review by ChatGPT and Human Experts," *Cureus* 15(4). (https://doi.org/10.7759/CUREUS.37281).

Wagner G., Prester, J. Mousavi, R., Lukyanenko R., and Paré G. (2026). Generative Artificial intelligence for literature reviews, ***Journal of Information Technology***. *Open access*

Thelwall, M., and Pardeep, S. (2022). "Scopus 1900–2020: Growth in Articles, Abstracts, Countries, Fields, and Journals," *Quantitative Science Studies* (3:1), pp. 37–50. (https://doi.org/10.1162/QSS_A_00177).

Theobald, E. J., Ettinger, A. K., Burgess, H. K., DeBey, L. B., Schmidt, N. R., Froehlich, H. E., Wagner, C., HilleRisLambers, J., Tewksbury, J., Harsch, M. A. (2015). "Global Change and Local Solutions: Tapping the Unrealized Potential of Citizen Science for Biodiversity Research," *Biological Conservation* (181), 236–244. (https://doi.org/10.1016/J.BIOCON.2014.10.021).

Thießen, F., D'Souza, J., and Stocker, M. (2023). *Probing Large Language Models for Scientific Synonyms*.

Tingelhoff, F., Brugger, M., and Leimeister, J. M. (2025). A guide for structured literature reviews in business research: The state-of-the-art and how to integrate generative artificial intelligence. *Journal of Information Technology*, 40(1), 77-99. (https://doi.org/10.1177/02683962241304105)

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). "Llama: Open and Efficient Foundation Language Models," *arXiv Preprint arXiv:2302.13971*. (https://doi.org/10.48550/ARXIV.2302.13971).

Tversky, A., and Kahneman, D. (1974). "Judgment Under Uncertainty: Heuristics and Biases: Biases in Judgments Reveal Some Heuristics of Thinking Under Uncertainty," *Science* 185(4157), 1124–1131. (https://doi.org/10.1126/SCIENCE.185.4157.1124).

Vaswani, A., et al. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 5998–6008.

van de Schoot, R., Bruin, J. de, Schram, R., Zahedi, P., Boer, J. de, Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., and Oberski, D. L. (2021). "An Open Source Machine Learning Framework for Efficient and Transparent Systematic Reviews," *Nature Machine Intelligence* 3(2), 125–133. (https://doi.org/10.1038/S42256-020-00287-7).

Van Den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural Discrete Representation Learning. In Advances in Neural Information Processing Systems (NeurIPS). arXiv:1711.00937

van Wee, B., and Banister, D. (2023). "Literature Review Papers: The Search and Selection Process," *Journal of Decision Systems*, pp. 1–7. (https://doi.org/10.1080/12460125.2023.2197703).

Vert, J.-P. (2023). "How Will Generative AI Disrupt Data Science in Drug Discovery?" *Nature Biotechnology* 41(6), 750–751. (https://doi.org/10.1038/S41587-023-01789-6).

Wagner, G., Empl, P., and Schryen, G. (2020). "Designing a Novel Strategy for Exploring Literature Corpora," in *European Conference on Information Systems*. (https://aisel.aisnet.org/ecis2020_rp/44).

Wagner, G., Lukyanenko, R., and Paré, G. (2022). "Artificial Intelligence and the Conduct of Literature Reviews," *Journal of Information Technology* 37(2), 209–226. (https://doi.org/10.1177/02683962211048201).

Wang, S., Scells, H., Koopman, B., and Zuccon, G. (2023). "Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search?" in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1426–1436. (https://doi.org/10.1145/3539618.3591703).

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *Proceedings of the 2019 International Conference on Learning Representations*.

Weber, R. (2024). "The Other Reviewer: RoboReviewer," *Journal of the Association for Information Systems* 25(1), 85–97. (https://doi.org/10.17705/1JAIS.00866).

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022a). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022b). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Advances*

Wagner G., Prester, J. Mousavi, R., Lukyanenko R., and Paré G. (2026). Generative Artificial intelligence for literature reviews, ***Journal of Information Technology***. *Open access*

in Neural Information Processing Systems (35), 24824–24837. (https://doi.org/10.48550/ARXIV.2201.11903).

Williamson, T. (2002). *Knowledge and Its Limits*, Oxford University Press, USA.

Xu, Xiaohan, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. (2024). "A survey on knowledge distillation of large language models," *arXiv preprint arXiv:2402.13116*.

Yan, C., Grabowska, M. E., Dickson, A. L., Li, B., Wen, Z., Roden, D. M., Michael Stein, C., Embí, P. J., Peterson, J. F., Feng, Q., Malin, B. A., and Wei, W.-Q. (2024). "Leveraging Generative AI to Prioritize Drug Repurposing Candidates for Alzheimer's Disease with Real-World Clinical Validation," *NPJ Digital Medicine* 7(1), 46. (https://doi.org/10.1038/S41746-024-01038-3).

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., and Chen, Y. (2023). "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," *arXiv Preprint arXiv:2309.01219*.

Zhang, R., Indulska, M., & Sadiq, S. (2019). Discovering data quality problems: the case of repurposed data. *Business & Information Systems Engineering 61*, 575-593.

Zhao, J., Huang, F., Lv, J., Duan, Y., Qin, Z., Li, G., and Tian, G. (2020). Do RNN and LSTM have long memory?. In *International Conference on Machine Learning* (pp. 11365-11375). PMLR.

Zhong, Q., Wang, K., Xu, Z., Liu, J., Ding, L., Du, B., and Tao, D. (2024). *Achieving >97*. (https://arxiv.org/abs/2404.14963).

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2020). "A Comprehensive Survey on Transfer Learning," *Proceedings of the IEEE* 109(1), 43–76. (https://doi.org/10.1109/JPROC.2020.3004555).

Zoph, B., Raffel, C., Schuurmans, D., Yogatama, D., Zhou, D., Metzler, D., Chi, E. H., Wei, J., Dean, J., Fedus, L. B., Bosma, M. P., Vinyals, O., Liang, P., Borgeaud, S., Hashimoto, T. B., and Tay, Y. (2022). "Emergent Abilities of Large Language Models," *Transactions on Machine Learning Research*. (https://doi.org/10.48550/ARXIV.2206.07682).

Zur Schlemmer, S. (2024). "Is It Possible for Artificial Intelligence to Undermine the Root of Science?" *Science Insights* 44(1), 1229–1234. (https://doi.org/10.15354/SI.24.RE881)

Online Supplementary Appendix: Evaluation

For the evaluation of the prompts by the author team, we purposely selected six papers from prominent information systems journals. We selected papers from different publishers (i.e., SAGE, Elsevier, Association for Information Systems, University of Minnesota, INFORMS, Taylor & Francis), with different writing styles (i.e., forward-looking, critical, integrative, formal, technical), and different methodologies (i.e., design science, theoretical literature review, qualitative case study, computational analysis, cross-sectional survey) to test the prompts against different texts. For each prompt, we used the PDF file of each of the six papers as the context. We then evaluated each prompt against the state-of-the-art GenAI models GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro. The pre-training knowledge cut-off dates were October 2023 for GPT-4o, April 2024 for Claude 3.5 Sonnet, and November 2023 for Gemini 1.5 Pro. For each evaluation, the assessments of success reflect the authors' qualitative judgments based on systematic evaluation, rather than formal empirical measures of success. A prompt was deemed successful if it produced a useful and accurate output that supported the intended literature review activity (as summarized in Table A1).

Table A1. Evaluation of prompts

Prompt	Paper	GPT-4o	Claude 3.5 Sonnet	Gemini 1.5 Pro
1. Prompt to identify prior review papers based on citation context	Brendel et al. (2021)	Yes (3/3)	Yes (3/3)	Yes (3/3)
	Sun & Gregor (2023)	Yes (1/1)	Yes (1/1)	Yes (1/1)
	Rinta-Kahila et al. (2023)	No (0/2)	No (0/2)	No (0/2)
	Kim et al. (2018)	No (1/2)	Yes (2/2)	No (1/2)
	Huber et al. (2017)	Yes (2/2)	Yes (2/2)	Yes (2/2)
	Tams & Dulipovici (2022)	Yes (1/1)	Yes (1/1)	Yes (1/1)
	Success rate [a]	**8 out of 11**	**9 out of 11**	**8 out of 11**
2. Concept definition prompt	Brendel et al. (2021)	No	No	Yes
	Sun & Gregor (2023)	Yes	Yes	Yes
	Rinta-Kahila et al. (2023)	No	Yes	Yes
	Kim et al. (2018)	No	No	No
	Huber et al. (2017)	Yes	Yes	Yes
	Tams & Dulipovici (2022)	Yes	No	Yes
	Success rate	**3 out of 6**	**3 out of 6**	**5 out of 6**

3. Prompt to assess the fit of a selected review type	Brendel et al. (2021)	No	No	No
	Sun & Gregor (2023)	No	No	No
	Rinta-Kahila et al. (2023)	Yes	Yes	Yes
	Kim et al. (2018)	No	No	No
	Huber et al. (2017)	No	No	No
	Tams & Dulipovici (2022)	Yes	No	No
	Success rate	**2 out of 6**	**1 out of 6**	**1 out of 6**
4. Language translation in the screening process	Brendel et al. (2021)	Yes	Yes	Yes
	Sun & Gregor (2023)	Yes	Yes	Yes
	Rinta-Kahila et al. (2023)	Yes	Yes	Yes
	Kim et al. (2018)	Yes	Yes	Yes
	Huber et al. (2017)	Yes	Yes	Yes
	Tams & Dulipovici (2022)	Yes	Yes	Yes
	Success rate	**6 out of 6**	**6 out of 6**	**6 out of 6**
5. Basic evaluation of the methodological approach	Brendel et al. (2021)	No	No	No
	Sun & Gregor (2023)	No	No	No
	Rinta-Kahila et al. (2023)	No	No	No
	Kim et al. (2018)	No	No	No
	Huber et al. (2017)	Yes	Yes	Yes
	Tams & Dulipovici (2022)	Yes	No	No
	Success rate	**2 out of 6**	**1 out of 6**	**1 out of 6**
6. Pseudocode for structured data extraction from tables	Brendel et al. (2021)	Yes	Yes	Yes
	Sun & Gregor (2023)	Yes	Yes	Yes
	Rinta-Kahila et al. (2023)	Yes	Yes	Yes
	Kim et al. (2018)	Yes	Yes	No
	Huber et al. (2017)	No	Yes	Yes
	Tams & Dulipovici (2022)	Yes	Yes	Yes
	Success rate	**5 out of 6**	**6 out of 6**	**5 out of 6**

Notes. [a] *The six example papers contain eleven references to prior review papers.*

To evaluate the retrieval-augmented generation (RAG) prompt (prompt 4), we used Elicit and Consensus. On both platforms, we assessed whether the first ten papers were relevant to the question, and we assessed the *adequacy of the summaries* based on four criteria: accuracy (i.e., whether the summary accurately captures the most important information from the source document), coherence (i.e., whether the content of the summary followed a logical flow and organization), consistency (i.e., whether the summary accurately reflected the facts, data, and conclusions presented in the original source document), and fluency (i.e., whether the summary was grammatically correct, well-phrased, and easily readable). In addition, we assessed the nature of the sources (see Table A2).

Table A2. Evaluation of prompt 4 (exploring prior research using a tabular overview)

Platform	**Paper**	**Relevance to the question**	**Adequacy of the summary**	**Nature of the source** [a]
Elicit	Yang et al. (2024b)	No	Adequate	Journal
	Lucas et al. (2024)	No	Adequate	Journal
	Sorin et al. (2023)	No	Adequate	Preprint
	Sirazitdinov et al. (2024)	No	Adequate	Conference
	Petersen (2011)	No	Adequate	Journal
	Wagner and Ruhe (2018)	No	Adequate	Preprint
	Elizalde and Bayona (2018)	No	Adequate	Conference
	Navarro-Cota et al. (2024)	No	Adequate	Journal
	Yang et al. (2024a)	Yes	Adequate	Preprint
	Patel et al. (2024)	Yes	Adequate	Preprint
	Summary	**2 out of 10**	**10 out of 10**	**4 Preprints, 2 Conference papers, 2 Journal papers**
Consensus	Zaman et al. (2019)	No	Adequate	Journal
	Durak et al. (2023)	No	Adequate	Journal
	Harrison and Rainer (1992)	No	Adequate	Journal
	Chapetta and Travassos (2020)	No	Adequate	Journal
	Graziotin and Abrahamsson (2014)	No	Adequate	Journal
	Bollati et al. (2023)	No	Adequate	Conference
	Orlando Lopez-Cruz et al. (2017)	No	Adequate	Conference
	Xinogalos et al. (2019)	No	Adequate	Journal
	Eloundou et al. (2023)	Yes	Adequate	Preprint
	Jeuring et al. (2023)	Yes	Adequate	Preprint
	Summary	**2 out of 10**	**10 out of 10**	**2 Preprints, 2 Conference papers, 4 Journal papers**

Notes. [a] All sources are open-access publications.

Wagner G., Prester, J. Mousavi, R., Lukyanenko R., and Paré G. (2026). Generative Artificial intelligence for literature reviews, ***Journal of Information Technology***. *Open access*

References

Bollati, V. A., Gaona, G., Lima, P. B., & Cuenca Pletsch, L. (2023). Software Development Teams: Factors Influencing their Productivity. *LACCEI*, 1(8).

Brendel, A. B., Lembcke, T. B., Muntermann, J., & Kolbe, L. M. (2021). Toward replication study types for design science research. *Journal of Information Technology*, 36(4), 340-355. (https://doi.org/10.1177/02683962211006429).

Chapetta, W. A., & Travassos, G. H. (2020). Towards an evidence-based theoretical framework on factors influencing the software development productivity. *Empirical Software Engineering*, 25, 3501-3543.

Durak, H. Y., Saritepeci, M., & Durak, A. (2023). Modeling of relationship of personal and affective variables with computational thinking and programming. *Technology, Knowledge and Learning*, 28(1), 165-184.

Elizalde, R., & Bayona, S. (2018). Interpersonal relationships, leadership and other soft skills in software development projects: a systematic review. *Trends and Advances in Information Systems and Technologies* 2(6), 3-15.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130.*

Graziotin, D., Wang, X., & Abrahamsson, P. (2014). Happy software developers solve problems better: psychological measurements in empirical software engineering. *PeerJ*, 2, e289.

Harrison, A. W., & Rainer Jr, R. K. (1992). The influence of individual differences on skill in end-user computing. *Journal of Management Information Systems*, 9(1), 93-111.

Wagner G., Prester, J. Mousavi, R., Lukyanenko R., and Paré G. (2026). Generative Artificial intelligence for literature reviews, ***Journal of Information Technology***. *Open access*

Huber, T. L., Kude, T., & Dibbern, J. (2017). Governance practices in platform ecosystems: Navigating tensions between cocreated value and governance costs. *Information Systems Research*, *28*(3), 563–584. https://doi.org/10.1287/isre.2017.0701

Jeuring, J., Groot, R., & Keuning, H. (2023). What skills do you need when developing software using chatgpt? (discussion paper). In *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research* (pp. 1-6).

López-Cruz, O., Mora, A. L., Sandoval-Parra, M., & Espejo-Gavilán, D. L. (2017). Teaching computer programing as knowledge transfer: Some impacts on software engineering productivity. In *Trends and Applications in Software Engineering: Proceedings of CIMPS 2016* 5 (pp. 145-154).

Lucas, H. C., Upperman, J. S., & Robinson, J. R. (2024). A systematic review of large language models and their implications in medical education. *Medical Education* 58(11), 1276-1285.

Navarro-Cota, C., Molina, A. I., Redondo, M. A., & Lacave, C. (2024). Individual differences in computer programming: a systematic review. *Behaviour & Information Technology*, 1-19.

Patel, H., Boucher, D., Fallahzadeh, E., Hassan, A. E., & Adams, B. (2024). A State-of-the-practice Release-readiness Checklist for Generative AI-based Software Products. *arXiv preprint arXiv:2403.18958.*

Petersen, K. (2011). Measuring and predicting software productivity: A systematic map and review. *Information and Software Technology*, 53(4), 317-343.

Rinta-Kahila, T., Penttinen, E., Salovaara, A., Soliman, W., & Ruissalo, J. (2023). The Vicious Circles of Skill Erosion: A Case Study of Cognitive Automation. *Journal of the Association for Information Systems*, *24*(5), 1378–1412. https://doi.org/10.17705/1jais.00829

Wagner G., Prester, J. Mousavi, R., Lukyanenko R., and Paré G. (2026). Generative Artificial intelligence for literature reviews, ***Journal of Information Technology***. *Open access*

Sirazitdinov, I., Succi, G., & Ivanov, V. (2020, December). A systematic literature review of studies related to mental activities of software developers. In *2020 International Conference Nonlinearity, Information and Robotics* (NIR) (pp. 1-11).

Sorin, V., Brin, D., Barash, Y., Konen, E., Charney, A., Nadkarni, G., & Klang, E. (2023). Large language models (LLMs) and empathy–a systematic review. *medRxiv Preprint*.

Sun, R., & Gregor, S. (2023). Reconceptualizing platforms in information systems research through the lens of service-dominant logic. *The Journal of Strategic Information Systems*, *32*(3), 101791. (https://doi.org/10.1016/j.jsis.2023.101791)

Tams, S., & Dulipovici, A. (2024). The creativity model of age and innovation with IT: Why older users are less innovative and what to do about it. *European Journal of Information Systems*, *33*(3), 287–314. https://doi.org/10.1080/0960085X.2022.2137064

Wagner, S., & Ruhe, M. (2018). A systematic review of productivity factors in software development. *arXiv preprint arXiv:1801.06475*.

Xinogalos, S., Satratzemi, M., Chatzigeorgiou, A., & Tsompanoudi, D. (2019). Factors affecting students' performance in distributed pair programming. *Journal of Educational Computing Research*, 57(2), 513-544.

Yang, Z., Sun, Z., Yue, T. Z., Devanbu, P., & Lo, D. (2024a). Robustness, security, privacy, explainability, efficiency, and usability of large language models for code. *arXiv preprint arXiv:2403.07506*.

Yang, X., Wang, Z., Wang, Q., Wei, K., Zhang, K., & Shi, J. (2024b). Large language models for automated Q&A involving legal documents: a survey on algorithms, frameworks and applications. *International Journal of Web Information Systems*, 20(4), 413-435.

Wagner G., Prester, J. Mousavi, R., Lukyanenko R., and Paré G. (2026). Generative Artificial intelligence for literature reviews, ***Journal of Information Technology***. *Open access*

Yongsuk Kim, Sirkka L. Jarvenpaa, & Gu, B. (2018). External Bridging and Internal Bonding: Unlocking the Generative Resources of Member Time and Attention Spent in Online Communities. *MIS Quarterly, 42*(1), 265–283. https://doi.org/10.25300/MISQ/2018/13278

Zaman, U., Jabbar, Z., Nawaz, S., & Abbas, M. (2019). Understanding the soft side of software projects: An empirical study on the interactive effects of social skills and political skills on complexity–performance relationship. *International Journal of Project Management*, 37 (3), 444-460.

Online Supplementary Appendix: Large Language Models (LLM) Architecture and Pre-training

Pre-trained language models are based on the Transformer architecture that was introduced by Vaswani et al. (2017). The original Transformer consists of two components: an encoder, which maps the input sequence into contextual representations, and a decoder, which generates an output sequence while attending to those representations. Following this breakthrough, three principal Transformer-based pre-training paradigms emerged:

1. Masked Language Modeling (MLM):

   Bidirectional encoders, typified by BERT (Devlin et al., 2019), train by randomly replacing input tokens with a *[MASK]* symbol and predicting each missing word from the surrounding context (Figure 1a). The bidirectional attention delivers high-quality sentence and document embeddings, which translate into superior performance on classification, natural-language inference, and named-entity recognition tasks. However, because tokens are predicted independently, MLMs typically struggle to produce coherent long-form text (Lewis et al., 2020a).

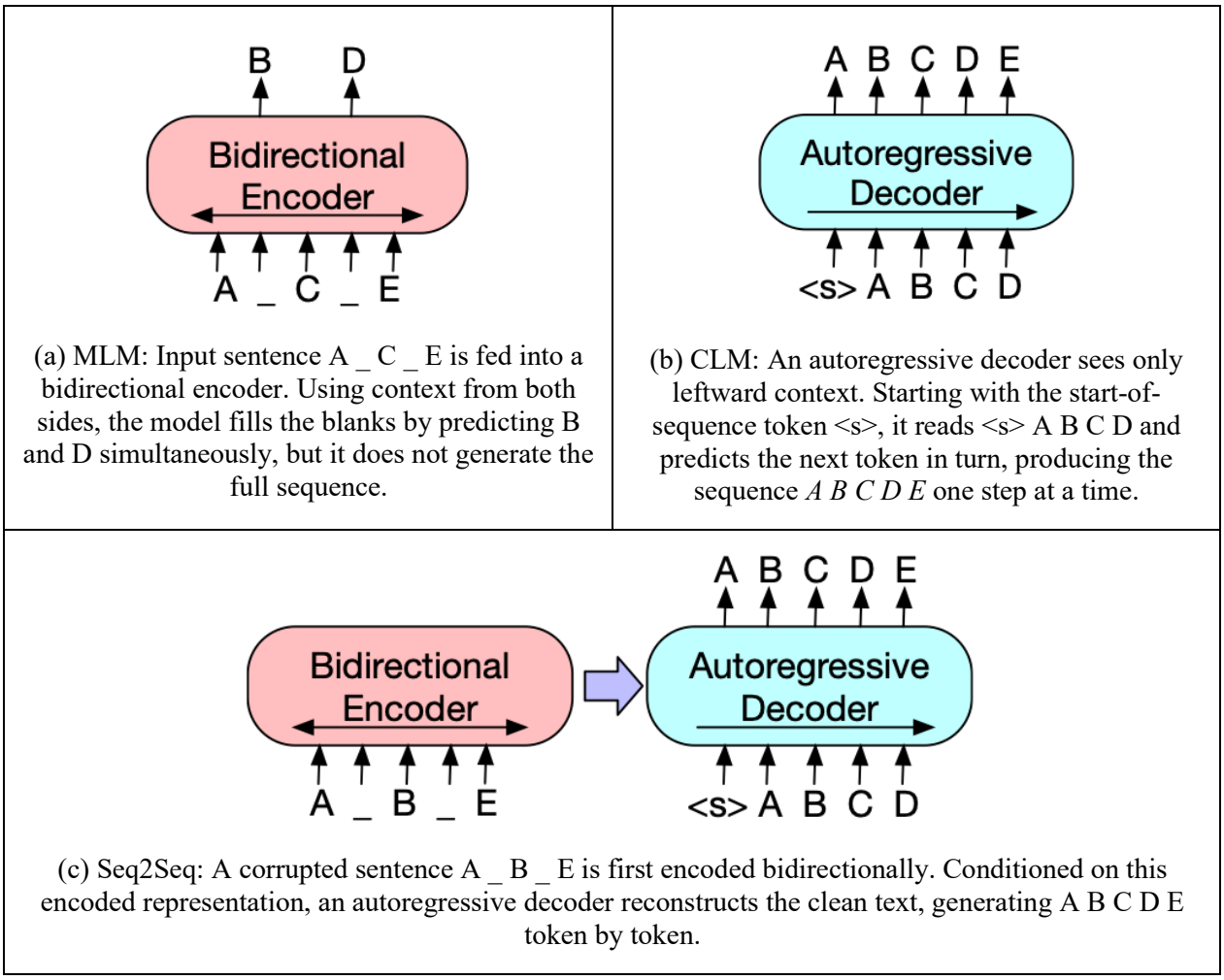
2. Causal Language Modeling (CLM):

   Autoregressive decoders, as used by the GPT family (Brown et al., 2020), predict the next token using only left-hand context (Figure 1b). This strict causality prevents target-token leakage and naturally aligns training with inference, enabling fluent left-to-right generation. CLMs therefore dominate applications that require open-ended text, dialogue, or code synthesis (Artetxe et al. 2022; Mousavi et al. forthcoming; Qiu et al. 2020).

3. Sequence-to-Sequence (Seq2Seq) Denoising:

   Models such as T5 (Raffel et al., 2020) and BART (Lewis et al., 2020b) corrupt the input (by span masking, deletion, or sentence permutation) then train a bidirectional encoder to read the noisy text and an autoregressive decoder to reconstruct the clean original (Figure 1c). This hybrid objective equips Seq2Seq models with both deep semantic understanding and strong generative

capability, making them suitable for translation, summarization, and question answering tasks (Lewis et al., 2020b).



(a) MLM: Input sentence A _ C _ E is fed into a bidirectional encoder. Using context from both sides, the model fills the blanks by predicting B and D simultaneously, but it does not generate the full sequence.

(b) CLM: An autoregressive decoder sees only leftward context. Starting with the start-of-sequence token <s>, it reads <s> A B C D and predicts the next token in turn, producing the sequence *A B C D E* one step at a time.

(c) Seq2Seq: A corrupted sentence A _ B _ E is first encoded bidirectionally. Conditioned on this encoded representation, an autoregressive decoder reconstructs the clean text, generating A B C D E token by token.

Figure 1. Canonical Pre-training Paradigms for Transformer-based Language Models (adapted from Lewis et al., 2020b)

Despite the benefits of the hybrid Seq2Seq architecture, researchers have gravitated toward the simpler decoder-only approach, which offers substantial advantages for building large, general-purpose models. In particular, contemporary LLMs created for open-ended text generation—including GPT (Brown et al., 2020), Llama (Touvron et al., 2023), and Mistral (Jiang et al., 2023)—employ decoder-only Transformer architectures (Figure 1b). They adopt this design because it (1) removes the need for cross-attention to an encoder, (2) reduces computational overhead during both training and inference, and (3) directly optimizes

the next-token prediction task that underlies all text-generation applications (Brown et al., 2020; Radford et al., 2018; Touvron et al., 2023). This design brought two main advantages:

- Scaling efficiency: When a decoder-only CLM is trained, it does exactly what it will do at inference (i.e., predict the next token from the left-hand context) so every token position is an identical, self-contained prediction task. That uniform task can be copied to many graphics processing units (GPUs). Each GPU core works on different slices of text without needing frequent coordination with others. This speeds up the training process and makes it more predictable (i.e., engineers can forecast throughput and training time with high confidence) (Hoffmann et al., 2022).
- Task universality: Once a decoder-only CLM is post-processed with techniques such as instruction tuning, retrieval-augmented generation, or chain-of-thought prompting, it can tackle analytic tasks that previously required bidirectional (MLM) or encoder–decoder (Seq2Seq) models, while still outperforming those architectures in extended, free-form text generation (Bai et al., 2023; Wei et al., 2022b). Although MLMs and Seq2Seq designs remain valuable for niche domains or compute-limited settings, decoder-only causal models now offer the most advantageous trade-off between versatility and scalability (Brown et al., 2020; Mousavi et al., forthcoming).

Parallel to the development of Transformer-based language models, other powerful generative techniques emerged for non-text modalities. Generative Adversarial Networks (GANs) introduced a novel adversarial training process to create realistic data (Goodfellow et al., 2014), while Variational Autoencoders (VAEs) learn a latent data distribution to synthesize new samples (Kingma & Welling, 2014). More recently, diffusion models have achieved state-of-the-art results in image synthesis by learning to reverse a gradual noising process (Ho et al., 2020; Sohl-Dickstein et al., 2015).

The concurrent advancements in text-based language models and modality-specific generative models created the conditions for an architectural convergence aimed at cross-modal reasoning rather than single-

modality generation. This trajectory culminated in Large Multimodal Models (LMMs), which explicitly integrate multiple data modalities within a unified framework. Models such as Google's Gemini (Gemini Team et al., 2024), Claude (Anthropic, 2024), DeepSeek (DeepSeek-AI et al., 2025), and OpenAI's GPT-4o (OpenAI, 2024) represent this next frontier, capable of processing and reasoning over text, images, audio, and video. Despite their expanded capabilities, these LMMs are not a departure from the CLM paradigm. At their core, they retain a CLM as the principal reasoning module, augmenting it with modality-specific encoders that translate non-text inputs into a form the CLM can process (Liu et al. 2023).

The key architectural advance is a token-level unification layer. Consider the visual pathway: an input image is first encoded by a vision network (such as a Vision Transformer) that produces a sequence of fixed-length embeddings (often called image tokens). Through a learned projection, these embeddings are mapped into the same semantic vector space as textual tokens and concatenated, typically as a prefix or interleaved sequence, with the user's text prompt. The resulting multimodal sequence is passed to the causal decoder, which applies its standard autoregressive, next-token objective; the model therefore generates text that is jointly conditioned on both linguistic context and the encoded visual content (Liu et al., 2023). By generalizing this mechanism to additional encoders (e.g., for audio or video), LMMs achieve coherent cross-modal reasoning without abandoning the computational advantages of the decoder-only Transformer backbone.

How Modern GenAI Models Generate Text

As discussed in the previous section, contemporary large language models intended for open-ended generation (e.g., GPT-4 and Llama 3.1) adopt a decoder-only CLM because it removes the computational overhead of cross-attention, scales linearly across GPUs, and aligns training with inference. Since our current work centers on text-generation tasks, we restrict the technical exposition to this dominant architecture and defer multimodal extensions to other work.

The text generation process, initiated by an input (e.g., a prompt), involves transforming this input into vector representations. These vectors are then processed by a decoder, which predicts the next word in the sequence. For example, given the input prompt "How is the food?", the decoder might predict "It" as the next word, which is then appended to the original input to create the input for predicting the following word (Figure 2). This iterative prediction generates a series of potential next words along with their probabilities, forming a probability distribution across the vocabulary. In a basic approach, called greedy sampling (Holtzman et al. 2020), the model always selects the word with the highest probability of occurring (as depicted Figure 2) but this can result in repetitive and predictable outputs. An alternative approach, called random sampling, adds some randomness to avoid always selecting the words with the highest probabilities. This randomness is regulated by two critical hyperparameters: Temperature and top-p. Temperature influences the degree of randomness in selecting words—lower values lead to more predictable text, while higher values encourage diversity. Top-p defines a threshold to select a subset of probable words, balancing coherence and variation in the output (see Table A1).

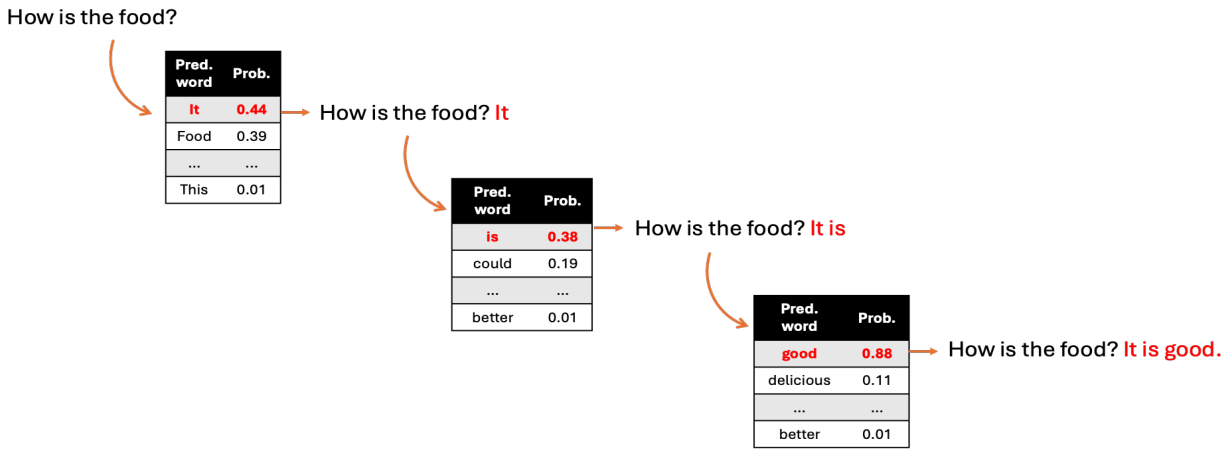


Figure 2. A simplified illustrative example of text generation in autoregressive LLMs using greedy sampling. The model selects the most probable word based on the patterns learned during pre-training.

Table 1. Greedy Sampling, Temperature, and Top-p

Strategy	Mechanism	Typical effect
Greedy sampling	At every step, the model simply chooses the single word with the highest probability.	Deterministic but often repetitive or overly cautious.
Temperature	Before choosing the next word, the model sharpens or flattens its probability distribution by dividing the scores by a temperature value. A low temperature (for example 0.3) makes the model pick high-probability words more often; a high temperature (for example 1.1) lets it consider less-likely words.	Lower temperatures yield safer, more predictable sentences; higher temperatures increase diversity and novelty but can introduce errors.
Top-p	The model first gathers the smallest set of candidate words whose combined probability reaches a chosen threshold p (commonly 0.8 or 0.9). It then picks randomly from just that shortlist instead of the whole vocabulary.	Removes extremely unlikely words while still allowing variation, striking a balance between coherence and novelty.

References

Anthropic AI (2024). The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1(1), 4.

Artetxe, M., Du, J., Goyal, N., Zettlemoyer, L., and Stoyanov, V. (2022). On the role of bidirectionality in language model pre-training. *arXiv preprint arXiv:2205.11726*. (https://doi.org/10.48550/arXiv.2205.11726)

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., et al. (2023). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*. (https://doi.org/10.48550/arXiv.2204.05862)

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). *Language Models Are Few-Shot Learners*. (https://arxiv.org/abs/2005.14165).

DeepSeek-AI (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. (https://arxiv.org/pdf/2501.12948)

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*

Wagner G., Prester, J. Mousavi, R., Lukyanenko R., and Paré G. (2026). Generative Artificial intelligence for literature reviews, ***Journal of Information Technology***. *Open access*

American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171-4186). (https://doi.org/10.18653/v1/N19-1423)

Gemini Team., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., et al. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*. (https://doi.org/10.48550/arXiv.2403.05530)

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems* 27, 2672-2680.

Ho, J., Jain, A. N., and Abeel, P. (2020). Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 6840–6851.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., et al. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*. (https://doi.org/10.48550/arXiv.2203.15556)

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*. (https://doi.org/10.48550/arXiv.1904.09751)

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Renard Lavaud, L., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., and El Sayed, W. (2023). Mistral 7B. *arXiv:2310.06825* (https://doi.org/10.48550/arXiv.2310.06825)

Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., and others. (2020a). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems* (33), 9459–9474. (https://doi.org/10.5555/3495724.3496517).

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., and others. (2020b). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and

Wagner G., Prester, J. Mousavi, R., Lukyanenko R., and Paré G. (2026). Generative Artificial intelligence for literature reviews, ***Journal of Information Technology***. *Open access*

Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871-7880.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 34892-34916.

Mousavi, R., Kitchens, B., Oliver, A., and Abbasi, A. (forthcoming). From Lexicons to Large Language Models: A Holistic Evaluation of Psychometric Text Analysis in Social Science Research. *Information Systems Research* (https://dx.doi.org/10.2139/ssrn.4776480)

OpenAI, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., et al. (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*. (https://doi.org/10.48550/arXiv.2410.21276)

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., and Huang, X. (2020). Pre-trained Models for Natural Language Processing: A Survey. *arXiv preprint arXiv:2003.08271*. (https://doi.org/10.1007/s11431-020-1647-3)

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Sohl-Dickstein, J., et al. (2015). Deep Unsupervised Learning using Nonequilibrium Thermodynamics. *Proceedings of the 32nd International Conference on Machine Learning*, 2256–2265.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). "Llama: Open and Efficient Foundation Language Models," *arXiv Preprint arXiv:2302.13971*. (https://doi.org/10.48550/ARXIV.2302.13971).

Vaswani, A., et al. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 5998–6008.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022b). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Advances in Neural Information Processing Systems* (35), 24824–24837. (https://doi.org/10.48550/ARXIV.2201.11903).

.