

# A Review on Scientific Knowledge Extraction using Large Language Models in Biomedical Sciences

Gabriel Lino Garcia<sup>1</sup> <sup>a</sup>, João Renato Ribeiro Manesco<sup>1</sup> <sup>b</sup> Pedro Henrique Paiola<sup>1</sup> <sup>c</sup>, Lucas Miranda<sup>2</sup>, Maria Paola de Salvo<sup>2</sup> and João Paulo Papa<sup>1</sup> <sup>d</sup>

<sup>1</sup>School of Sciences, São Paulo State University (UNESP), Bauru - SP, Brazil

<sup>2</sup>EasyTelling, São Paulo - SP, Brazil

{gabriel.lino, joao.rmanesco, pedro.paiola, joao.papa}@unesp.br  
{lucas, paola}@easytelling.com

**Keywords:** evidence synthesis, large language models, LLMs, biomedical, healthcare, literature-based discovery, knowledge extraction

**Abstract:** The rapid advancement of large language models (LLMs) has opened new boundaries in the extraction and synthesis of medical knowledge, particularly within evidence synthesis. This paper reviews the state-of-the-art applications of LLMs in the biomedical domain, exploring their effectiveness in automating complex tasks such as evidence synthesis and data extraction from a biomedical corpus of documents. While LLMs demonstrate remarkable potential, significant challenges remain, including issues related to hallucinations, contextual understanding, and the ability to generalize across diverse medical tasks. We highlight critical gaps in the current research literature, particularly the need for unified benchmarks to standardize evaluations and ensure reliability in real-world applications. In addition, we propose directions for future research, emphasizing the integration of state-of-the-art techniques such as retrieval-augmented generation (RAG) to enhance LLM performance in evidence synthesis. By addressing these challenges and utilizing the strengths of LLMs, we aim to improve access to medical literature and facilitate meaningful discoveries in healthcare.

## 1 INTRODUCTION

In recent years, the number of studies published in various fields has grown exponentially, with a wide range of applications spanning multiple disciplines (Bornmann et al., 2024). While this surge in scientific literature contributes to knowledge expansion, it poses a significant challenge for researchers and professionals attempting to stay informed of critical developments (Beller et al., 2013). This trend is particularly important in fields such as medicine, where the implications of new findings can have direct and life-altering impacts on patients' lives.

However, despite the abundance of information, the practical use of many of these studies remains limited (Meho, 2007). This is often due to barriers to accessing relevant and impactful research, particularly when important findings are hidden within the vast-

ness of the literature. The medical field exemplifies this challenge, as impactful discoveries that could improve treatment outcomes or advance medical science may go unnoticed simply because they are buried beneath a deluge of publications or presented in a highly technical manner.

Systematic reviews have long been regarded as a solution, providing comprehensive and structured analyses of the available literature. While effective, these reviews are exhaustive and frequently very technical, making them time-consuming to produce and difficult for non-experts to interpret. Consequently, they do not fully solve the issue of accessibility, especially for those who lack the specialized knowledge to extract meaningful insights from the data.

The advent of large language models (LLMs) offers a promising avenue for addressing this challenge. LLMs, powered by advancements in artificial intelligence, have the potential to assist in the navigation and synthesis of vast amounts of scientific literature. Indeed, some research has already begun to explore the potential of these models to automate aspects of the systematic review process, highlighting the pos-

<sup>a</sup>  <https://orcid.org/0000-0003-1236-7929>

<sup>b</sup>  <https://orcid.org/0000-0002-1617-5142>

<sup>c</sup>  <https://orcid.org/0000-0001-9093-535X>

<sup>d</sup>  <https://orcid.org/0000-0002-6494-7514>

sibility of their application in evidence synthesis and knowledge extraction (Alshami et al., 2023).

LLMs, such as GPT, are particularly well-suited to tasks involving knowledge extraction and summarization. They can process and synthesize vast amounts of text, distilling complex ideas into more accessible formats (Ignat et al., 2023). This capability makes them powerful tools for navigating large datasets and filtering out the most relevant information. Moreover, the ability of LLMs to generate concise and coherent summaries from complex data sets is already being harnessed in some preliminary applications within systematic review processes.

While early applications of LLMs in systematic reviews are promising, synthesizing scientific evidence for decision-making requires more than mere automation of reviews. Managers, healthcare professionals, and other non-technical stakeholders require tools to explore the literature and synthesize evidence meaningfully and effectively. This necessitates further investigation into how LLMs can be adapted as a tool for evidence synthesis for researchers and those involved in policy-making and management in fields like healthcare.

Given the potential of LLMs to address these challenges, this paper aims to provide a comprehensive review of current research exploring the use of LLMs in evidence synthesis, with a particular focus on their application in the medical field. We will examine existing methods, tools, and frameworks that leverage LLMs for scientific knowledge extraction and discuss how these approaches can help bridge the gap between vast amounts of data and practical, actionable insights.

Thus, we have proposed a review of scientific evidence synthesis using LLMs, particularly for the healthcare case. Among the contributions of our work are:

- We provide a comprehensive evaluation of the state-of-the-art LLMs used for medical evidence synthesis, focusing on their performance and challenges in tasks like clinical decision support and knowledge extraction.
- We identify key limitations, including issues with hallucinations, contextual understanding, and scalability, offering insights into the research gaps that need to be addressed.
- We propose the need for unified benchmarks and highlight future directions to enhance the effectiveness of LLMs in biomedical applications.

The remainder of this paper is structured as follows. In Section 2, we present the protocol used to conduct our review. Section 3 gives a background on

evidence synthesis based on scientific literature. Section 4 discusses the main works found in scientific evidence synthesis. Section 5 discusses the state of the art and current limitations in the literature, and Section 6 concludes the paper.

## 2 Review Methodology

This systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure a transparent and rigorous approach to evaluating the role of LLMs in the biomedical field. The main objective of this review is to provide a comprehensive assessment of the effectiveness, challenges, and methodologies of LLMs in the extraction and synthesis of medical knowledge. While knowledge extraction is a broader area of interest, this review specifically focuses on evidence synthesis, which aligns more closely with the complex task of integrating diverse medical information for clinical decision-making and research purposes.

The guiding research question of this review is: “How effective are LLMs in extracting and synthesizing structured and relevant knowledge from medical texts, and what are the key challenges and outcomes associated with their application in the medical field?” This question encompasses the exploration of the models most frequently used, how they are fine-tuned for biomedical applications, their performance compared to other knowledge extraction methods, and the specific biomedical domains that benefit most from LLM-based approaches.

The review adopted a search strategy across several academic and preprint databases, including ACM Digital Library, IEEE Xplore, Scopus, SOLO (including arXiv), and ScienceDirect. The search string was designed to cover the breadth of LLM applications in medical knowledge extraction, focusing specifically on evidence synthesis tasks. Studies were included if they focused on applying LLMs in medical settings, provided quantitative evaluations, and were published in English over the past five years.

### 2.1 Screening of Relevant Works

To enhance the screening process’s efficiency, we used a ChatGPT-4 LLM to generate summaries and provide context for each selected study. This approach allowed us to hasten the review of many articles while maintaining a high level of consistency in the analysis. The initial screening was guided by pre-defined inclusion and exclusion criteria, with priority given to studies that addressed LLM applications in

evidence synthesis. A diagram illustrating the screening and selection process, following PRISMA guidelines, is provided in Figure 1.

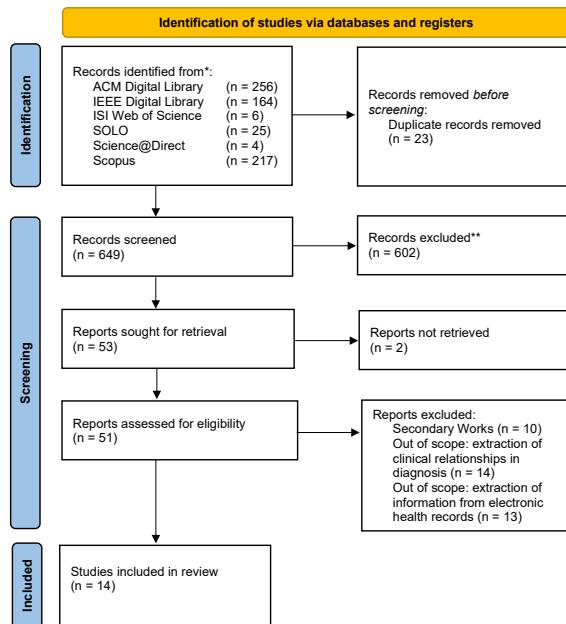


Figure 1: Illustration of the screening process, conducted in accordance with the PRISMA protocol, used to identify and select relevant studies for this review.

After the automated summarization phase, we manually validated the results to ensure that key insights, challenges, and contributions were accurately captured. This dual approach—combining automation with manual oversight—provided a robust method for managing the complexity of the literature while preserving the depth and rigor expected in systematic reviews.

For each included study, we extracted data on the type of LLM used, the medical task it was applied to, performance outcomes, and the challenges encountered. Particular attention was paid to how LLMs were adapted for biomedical knowledge extraction, the domains where they performed most effectively, and how they compared to other knowledge extraction techniques.

### 3 Background

Evidence synthesis refers to collecting, analyzing, and integrating information from the literature in a well-documented way to answer a pre-defined research question. One such form of evidence synthesis is through systematic reviews, which follow a rigorous search protocol to obtain all possible empiri-

cal evidence regarding a certain area of knowledge in a way that reduces bias and allows for replicability (Lasserson et al., 2019).

The objective of such systems is to support decision-makers by providing a reliable summary of research findings in the scope of the literature. However, constantly performing manual evidence synthesis whenever you need to make a decision can cause many constraints, especially since systematic reviews are highly complex and time-consuming, requiring strict adherence to rigorous protocols. This process often limits the decision maker’s ability to respond quickly to novel evidence (Singh, 2017).

Many approaches have been proposed to accelerate such processes using novel Natural Language Processing (NLP) techniques such as LLMs, mostly focusing on different aspects of the systematic review protocol, which can significantly reduce the manual burden and time required to complete a review (Bolanos et al., 2024). These tools, however, are still focused on specific parts of the review process, such as screening or data extraction. While they can still be relevant, preparing the complete review can still take time and hinder decision-makers productivity.

The use of LLMs to accelerate knowledge synthesis has been explored in other reviews (Bolanos et al., 2024; Burgard and Bittermann, 2023; de la Torre-López et al., 2023). However, these reviews primarily concentrate on streamlining the systematic review process. In contrast, our work focuses on techniques that use LLMs to perform evidence synthesis directly. In other words, those methods focus on compiling relevant literature in a more flexible, less constrained manner to provide quick, up-to-date knowledge, particularly within the healthcare domain, where timely access to synthesized evidence is crucial.

## 4 Results

This section synthesizes the selected studies that have applied various LLMs to medical evidence synthesis. The discussion is divided into two parts: first, we examine the use of LLMs in evidence synthesis as a broader concept, and second, we focus specifically on their applications within the medical field as reported in the literature.

### 4.1 Evidence Synthesis

Regarding evidence Synthesis using LLMs, a few works have been proposed more broadly to solve specific tasks in recent literature. One work (Khaira

et al., 2024) evaluates the efficacy of GPT-4 in performing systematic review tasks such as title/abstract screening, full-text review, and data extraction from various literature types and languages without human intervention. The study employed a "human-out-of-the-loop" approach, meaning that GPT-4 was tested independently without human intervention during extraction, focusing on the literature regarding parenting in protracted refugee situations. This work describes several challenges of current GPT models, indicating that the model's performance is heavily biased by the distribution of relevant and irrelevant studies, displaying a tendency of over-excluding studies.

In other words, this concept is expanded in less generic fields, such as government report generation (Gupta et al., 2024), in which Google's Gemini Pro and OpenAI's GPT-3.5 Turbo are used to improve data extraction, analysis, and visualization processes by reading all existing reports to identify graphs and charts needing updates, extracting data points from these graphs and synthesizing information in their knowledge repositories to update their report graphs. Although the method does not exactly synthesize information in textual form, it performs a Retrieval-Augmented Generation (RAG) method to synthesize information in graphic form.

The work of Yu et al. (2023) focuses on enhancing the extraction of training data from language models, specifically using the GPT-Neo 1.3B model in a two-stage pipeline: suffix generation, where various sampling strategies are employed, including top-k sampling, nucleus sampling, and typical sampling, to produce candidate suffixes based on a given prefix; and suffix ranking, in which perplexity and additional metrics like Zlib entropy are used to evaluate and select the most likely suffixes from the generated candidates. Although interesting for several aspects of knowledge extraction, this technique is still dependent on a large number of candidates that are ranked to identify a single successful instance, only working in limited scenarios.

In order to summarize the knowledge base, Discrete Text Summarization (DeTS) was proposed as a novel unsupervised method for discrete text summarization, in which grammatical sequence scoring and independent key points from large textual datasets are used for natural language inference. The technique consists of two main components: candidate extraction and candidate matching. In the candidate extraction phase, the algorithm utilizes Semantic Role Labeling (SRL) to segment comments into smaller pieces. The candidate matching phase aligns these key points with sentences from the entire corpus using NLI algorithms, determining whether the sentences

entail the key points based on their semantic similarity. Although the method was successful in a restricted scenario of anonymous comments, there is still a concern regarding bias in the employed dataset.

Focusing more on the extraction of information in scientific texts, Dagdelen et al. (2024) focuses on joint-named entity recognition and relation extraction, allowing the models to identify entities and their relationships within materials science literature. The model architecture is based on a sequence-to-sequence framework, where GPT-3 and LLaMa-2 models are fine-tuned on a relatively small dataset of 400-650 annotated text-extraction pairs. Given the dataset's limitations and the relationships' complexity, the model still struggles to perform complete evidence synthesis.

## 4.2 Synthesis in Biomedical Sciences

In the healthcare domain, which is the primary focus of our analysis, one study explores the application of LLMs, specifically GPT-3.5 and GPT-4, for Literature-Based Discovery to generate research hypotheses Nedbaylo and Hristovski (2024). The authors employ a bifurcated prompt engineering technique, where the original prompt is split into two parts and executed in separate chat windows. In the first segment, the model generates a list of disease characteristics without disclosing the disease name. In contrast, the second suggests potential interventions based on the factors identified in the initial segment. The study conducts a qualitative evaluation and concludes that GPT-based methods heavily depend on preexisting bibliographic databases, such as Medline, which restricts their applicability to fields with well-established datasets. Additionally, the study finds that outputs from GPT-4 often lack technical depth and novelty, and they frequently fail to provide adequate references for the generated information.

Tao et al. (2024) employs the OpenAI GPT-4 model to evaluate its performance in answering specific questions related to HIV drug resistance from published scientific papers. The researchers designed an automated pipeline that transformed the text of 60 selected HIV drug resistance papers into markdown format, excluding sections like the introduction and discussion to focus on the methods and results. By posing 60 questions in two modes, multiple-question mode (all questions presented simultaneously) and single-question mode (one question at a time), with and without an instruction sheet containing specialized knowledge, the authors were able to perform a quantitative evaluation of the task. Although the method obtained a fairly good mean ac-

curacy of 86.9%, some aspects of the technique could be improved, such as the fact that the instruction sheet was not effectively utilized by the model, by not improving the accuracy of the responses, in addition, the method struggled with specific queries that required making inferences, especially when dealing with implicit information in the texts. Worst of all, the GPT-4 model was more likely to provide false positive answers when questions were submitted individually than when they were submitted together, indicating the presence of bias when dealing with certain aspects of the literature.

Using the Claude 2 LLM, one study performs a proof-of-concept design to evaluate the performance of LLMs in interpreting randomized controlled trials (RCTs) Gartlehner et al. (2024). To do that, the researchers selected a convenience sample of 10 open-access RCTs and focused on extracting 16 distinct data elements, which included study identifiers, participant characteristics, and outcome data. The data extraction pipeline involves mostly prompt engineering, where the crafted prompts were iteratively tested and refined to optimize the model's output. Although the method obtains interesting results, several concerns should be considered, such as the fact that open RCTs were used, meaning they could be in the training set of Claude 2. The authors also reported two cases of hallucinations from answers that the model did not know how to reply to.

Focusing on information retrieval to support knowledge synthesis in biomedical documents, one work uses the RoBERTa language model to work with the COVID-19 Open Research Dataset Saxena et al. (2022). To do that, an architecture was developed consisting of three main components that extract relevant information: Paragraph Retrieval, Triplet Retrieval from Knowledge Graphs, and Complex Question Answering. The Paragraph Retrieval employs a hybrid approach combining lexical methods with semantic search for indexing and re-ranking results based on relevance. The Triplet Retrieval system extracts subject-relation-object triplets from the knowledge graph, allowing for faceted refinement of search results. Finally, the complex QA system utilizes a Multi-hop Dense Retriever to handle multi-hop questions, iteratively retrieving relevant passages and employing both extractive (RoBERTa) and generative (Fusion-in-Decoder) readers to generate answers. The model relies on an older approach for language models, which, although smaller, has more difficulty when dealing with generative language answers; the authors also report difficulty in dealing with the amount of unstructured textual data reported in the medical literature.

This trend of using encoder-based smaller language models was also observed in two other works in the literature. One of them works on a dataset of biomedical articles collected from PubMed, aiming to summarize biomedical literature by using an encoder-decoder model based on BioBERT Alambo et al. (2022) To achieve its purpose, the researchers created a framework based on two main components: an entity-driven knowledge retriever that extracts facts based on named entities identified in the source documents, and a knowledge-guided abstractive summarizer that generates summaries by attending to both the source article and the retrieved facts.

The other work focuses on semi-automate data extraction for systematic literature using a BERT model pre-trained on a corpus of 100,000 open-access clinical publications, the main objective being extracting clinical relationships from the corpus in order to perform literature filtering Panayi et al. (2023). Both works, although displaying interesting results, suffer from the difficulties of dealing with complex unstructured data; the BioBERT-based models, although displaying better results, still report hallucination, as there is an attempt to generate structured text.

Other works are more focused on extracting specific information from the biomedical literature. One such work, named ChIP-GPT, fine-tuned a LLaMa architecture on Sequence Read Archive information obtained from biomedical database records by extracting relevant sentences and summarizing biomedical records, intending to identify chromatin immunoprecipitation (ChIP) targets and cell lines in these records. A summarization step was implemented to manage the input length limitations of the model for longer records. This involved selecting informative sentences while discarding irrelevant details, ensuring the model could generate concise and accurate answers. The final results display good accuracy in the task even when compared with human experts. However, it was observed that this summarization step still had a problem discarding relevant information for the model, which impacted the final results.

Following the same concept, another work focuses on extracting relevant data from scientific literature for chemical risk assessment, focusing on Bisphenol A Sonnenburg et al. (2024). To achieve that, the authors created a dataset containing 78 selected publications, primarily extracting results sections to form prompt-completion pairs so that they could perform fine-tuning on an LLM from the GPT-3 family named Curie model. The Curie model reported remarkable results, especially compared to other ready-to-use models that were not fine-tuned, even if the authors still have some concerns regarding treating stud-

ies with more complex experimental designs.

Finally, a study that focuses on using a Retriever-Augmented Generation (RAG) model to extract and deliver accurate medical knowledge about diabetes and diabetic foot care, specifically tailored for laypersons with an eighth-grade literacy level, was proposed Mashatian et al. (2024). The authors use the GPT-4 model to formulate user-friendly responses based on the retrieved context, offering a zero-shot solution to the problem. A set of 295 articles was used to provide context and be evaluated across 175 questions on various diabetes-related topics. This work offers an interesting solution that does not require training the LLM and displays the results in natural language, even for people with lower literacy levels. Despite its advantages, the model is still evaluated on a very limited set of data, so further experiments are required to see how it performs with new works being published in the literature.

## 5 Comparison of Current Methods and Future Directions

In this section, we perform an aggregated analysis that compares datasets, performance metrics, identified limitations, and future trends suggested by each method to understand the limitations of current techniques and the future trends in the literature. Although the area lacks a complete benchmark for evaluating evidence synthesis, and each method does a different analysis, we still compare each technique by their main points and context in Table 1.

Existing methods that use language models for knowledge synthesis in biomedical tasks employ various strategies and applications, addressing several aspects of knowledge extraction, including literature-based discovery, data extraction, and question-answering. One of the key challenges in this area is to perform a comprehensive evaluation capable of comparing the effectiveness of each method and LLM by themselves. Currently, each method follows its own protocol, with distinct datasets, evaluation criteria, and tasks, making it difficult to draw direct comparisons between models or assess their generalizability to new problems.

In the reviewed studies, several models have been used for evidence syntheses, such as GPT-3.5, GPT-4, Claude 2, and even encoder-based methods, such as RoBERTa and BioBERT, on a wide range of tasks ranging from generating research hypotheses to answering specific scientific queries and extracting clinical trial data. For example, Nedbaylo and Hristovski (2024) focuses on literature-based discovery through

a GPT-4 LLM, while Tao et al. (2024) employs the same LLM in a very distinct task and context: processing structured biomedical data for HIV drug resistance analysis. Although both methods use the same model and follow similar strategies, it is difficult to conclude if the observed limitations come from the technique or the contextual data, as there is no shared information among them. Both studies also highlight considerable limitations of current LLMs in dealing with insights from the data, especially when the model needs to handle implicit information to generate connections and process novel information.

Another current problem in the literature is that most methods focus on using pre-established datasets, such as Medline in the study by Nedbaylo and Hristovski (2024) or the curated HIV drug resistance literature in Tao et al. (2024), to create a knowledge base for the LLM and fine-tune the model. In this case, it is difficult to establish the influence of each piece of data in the response, and it also makes dealing with the novel and growing literature even more difficult, as there is a need to re-train the model. As such, strategies based on RAG, such as the one proposed by Mashatian et al. (2024), appear as a more reliable solution to deal with new pieces of data and may pose the path forward for new techniques that may incorporate modern RAG approaches at their core.

In the study of Gartlehner et al. (2024) that uses a Claude 2 model for the extraction of data from RCTs, the authors raise two main concerns regarding this area: the growing risks of hallucination caused by difficulties in dealing with unstructured data in biomedical research, and the potential bias coming from training on open-access data, which the model may already have seen during training. In addition, without a common benchmark, it is difficult to determine how these models would fare on more diverse datasets or in more complex, less predictable environments.

Studies using smaller, encoder-based models like RoBERTa, BioBERT, and BERT also exhibit this challenge. While they are effective for specific tasks like entity extraction and information retrieval, their scalability and ability to generate new insights are limited. This is particularly evident in models designed for more specialized tasks, such as ChIP-GPT for chromatin immunoprecipitation data extraction and the fine-tuned Curie model for chemical risk assessment. These models achieve impressive results within their respective domains but lack the flexibility to be easily compared or adapted across different biomedical tasks without a unified evaluation standard.

The absence of a unified benchmark also extends

Table 1: Comparison of various methods used in biomedical literature for knowledge extraction and evidence synthesis.

Paper	Year	Model	Biomedical Context	Reported Results	Brief Summary
Alambo et al. (2022)	2022	BioBERT	Biomedical articles and facts from biomedical knowledge bases.	Precision: 55.07%, Recall 7.97%	The method uses integrates named entities and facts from biomedical knowledge bases into transformer-based models to enhance the factual consistency of generated summaries in biomedical literature.
Saxena et al. (2022)	2022	RoBERTa + Fusion-in-Decoder	Unstructured text data from documents, articles, biomedical journals, and structured data like tables and electronic health records.	Precision: 81%, NDCG: 72%	A framework for complex information retrieval from biomedical documents is developed, utilizing paragraph retrieval, triplet retrieval from knowledge graphs, and complex question answering.
Panayi et al. (2023)	2023	BERT	Metrics for progression-free survival, patient age, treatment arm dosage, and eGFR measurements extracted from systematic literature reviews in oncology and Fabry disease.	F1-score: 73%	A machine learning approach using pre-trained BERT and CRF models was employed to automate data extraction from systematic literature reviews by identifying biomedical entities and their relations.
Sonnenburg et al. (2024)	2024	GPT-3	NAM-based toxicity data on Bisphenol A (BPA) extracted from scientific publications	Precision: 25%, Recall: 23%, F1-score: 50%	A fine-tuned version of GPT-3 was employed to extract and structure data from scientific publications for risk assessment, specifically targeting toxicological properties of bisphenol A (BPA).
Nedbaylo and Hristovski (2024)	2024	GPT-3.5 and GPT-4	Scientific papers and literature related to biomedical concepts	Only a qualitative evaluation was performed in this paper	A literature-based discovery technique based on the generation of research hypotheses via bifurcated prompt engineering.
Cinquin (2024)	2024	ChIP-GPT (LLaMa)	Metadata from Chromatin Immunoprecipitation Sequencing (ChIP-Seq) experiments.	Accuracy: 90%	A LLaMA-based model is fine-tuned to extract metadata from biomedical database records, targeting chromatin immunoprecipitation (ChIP) data.
Gartlehner et al. (2024)	2024	Claude 2	The research paper focuses on extracting several types of biomedical data elements from published studies, including study identifiers, characteristics of study participants, and outcome data.	Accuracy: 96.3%, F1-score: 98%	The study uses the Claude 2 LLM to extract several types of biomedical elements from published studies in medical literature.
Mashatian et al. (2024)	2024	GPT-4	Scientific papers related to diabetes and diabetic foot care.	Accuracy: 98%	A RAG model was developed to extract accurate medical knowledge about diabetes and diabetic foot care for laypersons with an eighth-grade literacy level, using prompt engineering.
Tao et al. (2024)	2024	GPT-4	HIV genetic sequence data, antiviral treatment histories, and the effects of mutations on susceptibility to antiviral drugs from papers on HIV drug resistance (HIVDR).	Accuracy: 86.9%, Recall 72.5%, Precision 87.4%	An automated pipeline using GPT-4 was created to evaluate the accuracy of responses to queries about published papers on HIV drug resistance.

to the issue of hallucinations and false positives, which have been identified across multiple studies. For example, the GPT-4 model in Tao et al. (2024)'s study showed a higher tendency for false positives in single-question mode, raising concerns about the consistency and reliability of LLMs when applied to individual biomedical queries. Likewise, the hallucinations observed in Claude 2's responses in Gartlehner et al. (2024)'s work underlines the need for more robust evaluation methods to assess generated content's accuracy and factual correctness.

As we can observe from the current state of the field, developing a unified benchmark to standardize the evaluation of LLMs across evidence synthesis in diverse biomedical tasks is crucial to advance the field. This benchmark must evaluate the model's accuracy, contextual comprehension, insights capability, generalizability, and the capacity to process unstructured data commonly found in this area. Furthermore, recent advancements in LLM retrieval techniques, such as RAG techniques, could significantly enhance evidence synthesis, as these techniques not only improve the quality of generated responses but also allow for the models to deal with the novel incoming literature, on top of enabling the retrieval of the relevant sources, ensuring that LLMs contribute

to evidence synthesis in a transparent and verifiable manner.

## 6 Conclusion

This review aims to highlight the increasing role of natural language processing and LLMs in biomedical research, particularly for automating tasks like literature-based discovery, clinical trial data extraction, and answering complex scientific questions. Although these models show great promise, they also reveal significant shortcomings. Many rely on preexisting datasets and struggle with more nuanced tasks such as making inferences, understanding context, and dealing with implicit information. Issues like hallucinations and false positives in several studies further underscore the need for stronger fact-checking and better handling of contextual information.

A major challenge is the lack of a unified benchmark across various tasks. The current variety of datasets and evaluation metrics makes it difficult to compare models directly or evaluate how well they generalize to new problems. Fine-tuned models like ChIP-GPT and Curie perform well in narrow areas but are less proven in broader contexts. Similarly,

older encoder-based models such as RoBERTa and BioBERT have scalability limitations and are difficult to manage the complexity of unstructured medical literature.

Looking ahead, efforts should focus on improving models' ability to grasp context, addressing the problem of hallucinations, and creating unified benchmarks to standardize evaluation across tasks. Integrating techniques like knowledge graphs, multi-hop reasoning, and retrieval-augmented generation (RAG) could help close the current performance gaps, especially for complex evidence synthesis. While LLMs hold great potential in healthcare, they still require improvements in architecture, evaluation methods, and unstructured data handling to fully realize their transformative potential in biomedical research and automated evidence synthesis.

## ACKNOWLEDGEMENTS

This study was funded by the São Paulo Research Foundation (FAPESP) grants 2013/07375 – 0, 2019/07665 – 4, 2023/14427 – 8, 2024/00789 – 8, and 2024/01336 – 7 as well as the National Council for Scientific and Technological Development grants 308529/2021 – 9 and 400756/2024 – 2.

## REFERENCES

- Alambo, A., Banerjee, T., Thirunarayan, K., and Raymer, M. (2022). Entity-driven fact-aware abstractive summarization of biomedical literature. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 613–620. IEEE.
- Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A. 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.
- Beller, E. M., Chen, J. K.-H., Wang, U. L.-H., and Glasziou, P. P. (2013). Are systematic reviews up-to-date at the time of publication? *Systematic reviews*, 2:1–6.
- Bolanos, F., Salatino, A., Osborne, F., and Motta, E. (2024). Artificial intelligence for literature reviews: Opportunities and challenges. *arXiv preprint arXiv:2402.08565*.
- Bornmann, L., Haunschild, R., and Mutz, R. (2024). Accelerating scientific progress with preprints. *Nature Computational Science*, 4(5):311–311.
- Burgard, T. and Bittermann, A. (2023). Reducing literature screening workload with machine learning. *Zeitschrift für Psychologie*.
- Cinquin, O. (2024). Chip-gpt: a managed large language model for robust data extraction from biomedical database records. *Briefings in bioinformatics*, 25(2):bbad535.
- 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.
- de la Torre-López, J., Ramírez, A., and Romero, J. R. (2023). Artificial intelligence to automate the systematic review of scientific literature. *Computing*, 105(10):2171–2194.
- Gartlehner, G., Kahwati, L., Hilscher, R., Thomas, I., Kugley, S., Crotty, K., Viswanathan, M., Nussbaumer-Streit, B., Booth, G., Erskine, N., et al. (2024). Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Research Synthesis Methods*.
- Gupta, R., Pandey, G., and Pal, S. K. (2024). Automating government report generation: A generative ai approach for efficient data extraction, analysis, and visualization. *Digital Government: Research and Practice*.
- Ignat, O., Jin, Z., Abzaliev, A., Biester, L., Castro, S., Deng, N., Gao, X., Gunal, A., He, J., Kazemi, A., et al. (2023). Has it all been solved? open nlp research questions not solved by large language models. *arXiv preprint arXiv:2305.12544*.
- Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., and Hadfield, K. (2024). Can large language models replace humans in systematic reviews? evaluating gpt-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods*.
- Lasserson, T. J., Thomas, J., and Higgins, J. P. (2019). Starting a review. *Cochrane handbook for systematic reviews of interventions*, pages 1–12.
- Mashatian, S., Armstrong, D. G., Ritter, A., Robbins, J., Aziz, S., Alenabi, I., Huo, M., Anand, T., and Tavakolian, K. (2024). Building trustworthy generative artificial intelligence for diabetes care and limb preservation: A medical knowledge extraction case. *Journal of Diabetes Science and Technology*, page 19322968241253568.
- Meho, L. I. (2007). The rise and rise of citation analysis. *Physics World*, 20(1):32.
- Nedbaylo, A. and Hristovski, D. (2024). Implementing literature-based discovery (lbd) with chatgpt. In *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, pages 120–125. IEEE.
- Panayi, A., Ward, K., Benhadji-Schaff, A., Ibanez-Lopez, A. S., Xia, A., and Barzilay, R. (2023). Evaluation of a prototype machine learning tool to semi-automate data extraction for systematic literature reviews. *Systematic Reviews*, 12(1):187.

- Saxena, S., Sangani, R., Prasad, S., Kumar, S., Athale, M., Awhad, R., and Vaddina, V. (2022). Large-scale knowledge synthesis and complex information retrieval from biomedical documents. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 2364–2369. IEEE.
- Singh, S. (2017). How to conduct and interpret systematic reviews and meta-analyses. *Clinical and translational gastroenterology*, 8(5):e93.
- Sonnenburg, A., van der Lugt, B., Rehn, J., Wittkowski, P., Bech, K., Padberg, F., Eleftheriadou, D., Dobrikov, T., Bouwmeester, H., Mereu, C., et al. (2024). Artificial intelligence-based data extraction for next generation risk assessment: Is fine-tuning of a large language model worth the effort? *Toxicology*, 508:153933.
- Tao, K., Osman, Z. A., Tzou, P. L., Rhee, S.-Y., Ahluwalia, V., and Shafer, R. W. (2024). Gpt-4 performance on querying scientific publications: reproducibility, accuracy, and impact of an instruction sheet. *BMC Medical Research Methodology*, 24(1):139.
- Yu, W., Pang, T., Liu, Q., Du, C., Kang, B., Huang, Y., Lin, M., and Yan, S. (2023). Bag of tricks for training data extraction from language models. In *International Conference on Machine Learning*, pages 40306–40320. PMLR.