ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

ISPOR Report

A Taxonomy of Generative Artificial Intelligence in Health Economics and Outcomes Research: An ISPOR Working Group Report

Rachael L. Fleurence, PhD, Xiaoyan Wang, PhD, Jiang Bian, PhD, Mitchell K. Higashi, PhD, Turgay Ayer, PhD, Hua Xu, PhD, Dalia Dawoud, PhD, Jagpreet Chhatwal, PhD, on behalf of the ISPOR Working Group on Generative AI

ABSTRACT

Objectives: This article presents a taxonomy of generative artificial intelligence (AI) for health economics and outcomes research (HEOR), explores emerging applications, outlines methods to improve the accuracy and reliability of AI-generated outputs, and describes current limitations.

Methods: Foundational generative AI concepts are defined, and current HEOR applications are highlighted, including for systematic literature reviews, health economic modeling, real-world evidence generation, and dossier development. Techniques such as prompt engineering (eg, zero-shot, few-shot, chain-of-thought, and persona pattern prompting), retrieval-augmented generation, model fine-tuning, domain-specific models, and the use of agents are introduced to enhance AI performance. Limitations associated with the use of generative AI foundation models are described.

Results: Generative AI demonstrates significant potential in HEOR, offering enhanced efficiency, productivity, and innovative solutions to complex challenges. Although foundation models show promise in automating complex tasks, challenges persist in scientific accuracy and reproducibility, bias and fairness, and operational deployment. Strategies to address these issues and improve AI accuracy are discussed.

Conclusions: Generative AI has the potential to transform HEOR by improving efficiency and accuracy across diverse applications. However, realizing this potential requires building HEOR expertise and addressing the limitations of current AI technologies. Ongoing research and innovation will be key to shaping AI's future role in our field.

Keywords: artificial intelligence, economic models, generative AI, large language models, systematic reviews.

VALUE HEALTH. 2025; 28(11):1601–1610

Highlights

- This article introduces how generative AI can be applied in health economics and outcomes research (HEOR), presenting a taxonomy of terms, applications, and tools, along with methods to improve accuracy and reliability. Key challenges—scientific validity and reliability, bias and fairness, and operational deployment—must be addressed before AI can be fully integrated into HEOR.
- Generative AI shows promise in automating HEOR tasks, such as systematic reviews and economic modeling, but is not yet reliable for autonomous use. Techniques such as prompt engineering and retrieval-augmented generation can improve accuracy and dependability.
- Generative AI can improve HEOR processes but should augment rather than replace human expertise in the near term. HEOR professionals and healthcare decision makers should adopt generative AI tools, with strong checks and balances.

Introduction

Generative artificial intelligence (AI) is affecting multiple areas in science and medicine, including in health economics and outcomes research (HEOR).[1-3] The field of AI has been exploring ways to use machine intelligence to augment human endeavors since the 1950s.[4] By the 1990s, machine learning (ML) techniques were advancing pattern recognition and decision-making processes.[3] By the 2000s, researchers had developed deep learning models based on neural networks, enabling a wide range of complex applications from image recognition to natural language processing (NLP).[3,5] In 2021, a breakthrough in structural biology occurred when AlphaFold, a neural network-based deep learning program created by DeepMind, accurately predicted protein folding, significantly accelerating the process of drug discovery.[6,7] The scientists leading this effort were awarded a Nobel Prize in Chemistry in October 2024.[8]

In the past few years, foundation models (FMs)—large-scale AI systems trained on extensive, unlabeled data sets through self-supervised learning—have emerged as transformative tools.[9] These models represent a significant shift in healthcare AI, transitioning from task-specific, single-purpose models to more versatile and adaptable generalist AI systems for medical applications.[10,11] A major paradigm shift occurred in November 2022 with the launch of OpenAI's ChatGPT, a type of generative AI that produces text, images, or other content based on input prompts.[12,13] Large language models (LLMs), a key technology behind FMs, can recognize, summarize, and produce coherent and contextually relevant texts.[2] In recent years, several major FMs have emerged, including OpenAI's GPT models, Google's Gemini, Anthropic's Claude, and Meta's Llama.[14]

FMs have the potential to drive innovation in health technology assessment (HTA) domains, such as systematic literature reviews (SLRs), evidence synthesis, health economic modeling, real-world evidence (RWE) generation, and value dossier development.[3] These models can streamline research processes and significantly enhance productivity. With these early applications

Box 1
Key definitions.

Artificial intelligence (AI)
Artificial intelligence refers to the ability of machines to perform tasks that typically require human intelligence, such as pattern recognition, language understanding, reasoning, and decision making.[4] In HEOR, AI is increasingly used to automate complex and time-intensive tasks, including data extraction, statistical analysis, evidence synthesis, and predictive modeling, offering opportunities to enhance efficiency and accuracy in research.

Machine learning (ML)
Machine learning, an important subset of AI, allows algorithms to learn from data to perform tasks without explicit programming.[4,17] ML enables computers to adapt and improve their ability to solve problems over time. ML began to take root in the 1990s with the development of groundbreaking techniques, such as Support Vector Machines (SVM) and Random Forests.[18,19] ML has been used in HEOR to assist with various research tasks.[20]

Generative AI
Generative AI represents a more advanced application of AI, capable of creating new content, synthesizing data, and providing innovative solutions to complex problems.[3] Generative AI models, particularly FMs, can process and produce natural-language text, perform tasks that require reasoning, generate computer code, summarize research findings, draft reports, and much more.[2,12,14] These models are continuously improving and can be deployed to assist in various HEOR tasks, potentially providing greater accuracy and efficiency in research.[3]

Foundation models (FMs)
FMs are large-scale AI systems trained on massive datasets using self-supervised learning, enabling them to generalize across a wide range of tasks.[2,9,12] They are central to generative AI, offering versatility by performing well across multiple domains with minimal fine-tuning. Examples include GPT-4, Gemini, Claude, and LLaMA, used in applications ranging from text generation to image and video analysis.[14] Their adaptability makes them ideal for tasks such as information extraction, language understanding, text summarization, and visual data interpretation.

Large language models (LLMs)
LLMs are a specialized type of FM designed for processing and generating human language.[14] Trained on extensive text corpora, LLMs excel at tasks such as summarization, question answering, and content generation. Examples such as GPT-4, Claude, and LLaMA demonstrate competitive performance across diverse language tasks, often rivaling task-specific models with minimal fine-tuning.[21] Although all LLMs are FMs, they focus exclusively on language-based applications.

in HEOR emerging, HTA agencies are in the process of developing guidelines for the use of generative AI in submissions.[3,15] For instance, the National Institute for Health and Care Excellence in the United Kingdom has issued a statement of intent[16] to describe their approach to AI in general, as well as a position statement that covers the principles that should be adhered to when generative AI is used in submissions.[15]

This manuscript is intended to serve as a resource for HEOR professionals exploring the rapidly evolving field of generative AI and its impact on HEOR workflows. As adoption grows, understanding key terms, techniques, and applications is crucial. In particular, the article introduces fundamental concepts, such as AI, machine learning, generative AI, and FMs (defined in Box 1, illustrated in Fig. 1), highlights emerging HEOR applications, and reviews techniques to optimize generative AI use. A glossary is provided to explain key terms (Glossary). It also examines current limitations, offering a balanced view of the opportunities and risks associated with the use of FMs in the field of HEOR.

This article is not a comprehensive review of all HEOR applications or tools but intends to serve as a starting point for understanding these technologies. It offers foundational knowledge and practical insights for assessing their relevance in HEOR. Designed for a broad audience, it provides accessible explanations for beginners and advanced insights into techniques and limitations for more experienced readers exploring AI integration into HEOR workflows.

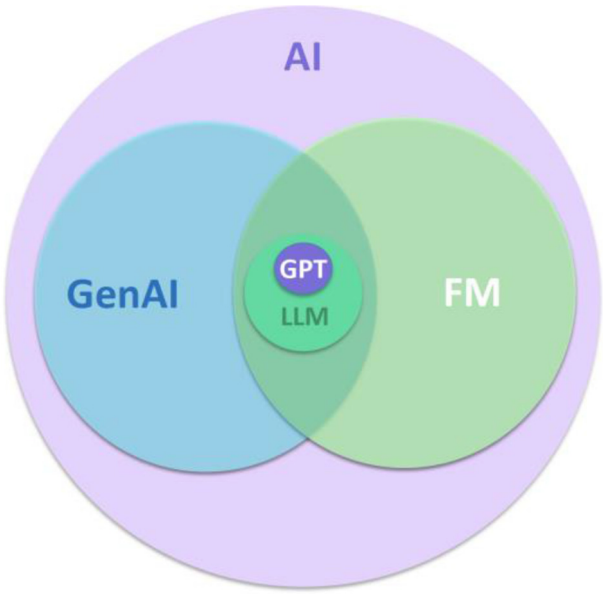
Examples of Applications of Generative AI in HEOR Use Cases

Applications using generative AI in HEOR are rapidly evolving, with use cases in SLRs, evidence synthesis, economic modeling, RWE generation, and dossier development.[3] Table 1 describes the diverse applications of generative AI in HEOR showcasing its potential to streamline and enhance key processes in the field. The following section highlights key applications and associated challenges, drawing on the authors' experiences and a targeted literature review. Although not exhaustive, it offers illustrative examples to guide HEOR professionals and showcase the diverse potential of generative AI.

Applications of Generative AI to SLRs and Network Meta-Analyses

An early exploration of generative AI and FMs has concentrated on SLRs, a critical process for evidence synthesis in health research.[3] SLRs are time consuming and labor intensive, requiring detailed and careful screening of abstracts and full-text articles, bias assessment, and precise and sometimes extensive data extraction and may include quantitative meta-analyses involving a vast number of studies.[22,23] FMs can streamline this process in a range of SLR tasks. Specifically, FMs can assist in developing the literature search strategies and screening abstracts and full-text articles for inclusion and exclusion using predefined criteria.[24-27] For example, a recent study demonstrated high accuracy using a GPT-4-based reviewer in PRISMA-based medical systematic literature reviews.[28] FMs can also provide reasoning for excluding certain abstracts and full-text articles.[29] They can assist with bias assessment by applying a list of questions to full-text articles.[30,31] These models can extract structured data from unstructured text from research articles.[32] They can be trained to identify key data points, such as population characteristics, interventions, comparators, and outcomes, improving the speed of data extraction.[33-35] FMs can also generate code for running meta-analyses in R and Python and other programming languages.[35] Additionally, they can generate concise summaries of the included studies, helping researchers synthesize findings more efficiently.[35] Network meta-

Figure 1. Relationship between AI, Gen AI, Foundation Models and LLMs.



analyses are widely used in HEOR to synthesize evidence from multiple studies and compare the relative effectiveness of interventions.[36] Recent advancements demonstrate the potential of FMs to streamline quantitative evidence synthesis by automating tasks such as extracting relevant study parameters, standardizing data inputs, and generating interpretable results.[35,37] For example, Reason et al showed that GPT-4 was able to replicate data extraction in 4 network meta-analyses with an accuracy exceeding 99%.[35] More generally, the application of FMs in automating statistical analyses represents a transformative opportunity for HEOR.[38,39]

FMs show promise in automating SLR tasks but still face limitations. There have been reports of hallucinations, in which FMs generate plausible but incorrect mesh terms of fabricated citations.[40]

Table 1. Applications of generative AI to areas of HEOR.

Area of HEOR	Potential and actual areas of use
Systematic literature reviews	Search strategy, abstract screening, full-text screening, bias assessment, data extraction, meta-analyses, and report writing.
Economic modeling	Economic model literature summarization, model conceptualization, parameter generation, code generation, structural uncertainty analysis, and report writing.
Real-world evidence	Data extraction, information retrieval, transformation of unstructured data to structured data, and integration of multimodal data.
Dossier development	Report writing in different styles and formats.

AI indicates artificial intelligence; HEOR, health economics and outcomes research.

Additionally abstract disposition and data extraction is not consistently accurate and require manual validation.[30] Although some studies demonstrate FM performance comparable to human tasks, this is not consistently reliable, emphasizing the need for human oversight.[40] Until robust methods and standards for evaluating generative AI outputs are established, manual verification remains essential.[3,15]

Applications of Generative AI to Health Economic Modeling

FMs can be used for a range of different applications in health economic modeling and have the potential to transform how models are conceptualized, developed, and utilized.[3] FMs can efficiently summarize existing economic models, providing a rapid synthesis of their methodologies and outcomes.[41] This capability is helpful for model parameterization and data extraction, in which FMs can expedite the identification and integration of relevant data points.[42] FMs can assist with the creation of new or de novo health economic models by leveraging extensive amounts of existing literature and data.[43] For example, one study demonstrated that a FM could replicate a 3-state partition model for non-small cell lung cancer and renal cell carcinoma.[44] Another proof-of-concept study showed that an FM could fully replicate a published simple health economic model evaluating the cost-effectiveness of combination therapy for HIV infection, including extraction of model structure, parameters identifications, code development, and results evaluation.[45]

FMs have the potential to assist with various higher complexity tasks in model development, although we are not aware of any published studies in this area at the time of writing. For instance, FMs could aid in validating existing models by cross-verifying assumptions and outputs against new data or parallel models. They might also support adapting models to different geographic or demographic contexts, improving their accuracy and applicability. Additionally, FMs could streamline platform transitions, such as converting models from Excel to R Shiny, by automating and error-checking the process. One of the more resource-intensive applications is conducting structural uncertainty analysis.[46] Here FMs could automate parts of the workflow, significantly reducing labor and time requirements. These current and potential applications highlight the ability of FMs to accelerate health economic modeling. However, human expertise and oversight remain essential because standards for ensuring the accuracy and reliability of AI-generated models are still evolving.[3,15,47,48]

Applications of Generative AI to RWE Generation

Generative AI and FMs have the potential to assist in generating RWE by improving efficiency, accuracy, and scale of real-world data that might be available for research. Only a small portion of electronic health records (EHR) data is structured and in a format that lends itself to statistical analysis with minimal processing.[49] Much detailed patient information is embedded in clinical documents and reports, which are in unstructured text. This unstructured data requires additional processing before it can be included in analytical data sets. For decades, advances in AI, particularly in NLP, have notably accelerated data extraction, information retrieval, and summarization for RWE generation.[50] Emerging applications include deploying generative AI tools to extract information from unstructured EHRs with the potential to accelerate their integration of these data into analyzable data sets, reduce manual efforts, and minimize human errors.[21,51] For example, FMs were shown to be successful in extracting biomarker testing details from EHR documents.[52] Some studies have questioned the accuracy of FMs in their ability to map descriptive text to medical codes, with one study finding that accuracy remains below 50%.[53] However, this study did

not use any fine-tuning, which may have improved the results. Approaches to improve the accuracy of mapping unstructured to structured text include the use of specialized models, such as GatorTron, NYUTron, and Me LLaMA, which are trained using large clinical texts,[54–57] and improving the prompts provided to GPT-3.5 and GPT-4.[21]

Generative AI tools show significant promise in enhancing RWE generation by streamlining the processing of unstructured data and integrating diverse data sources. By integrating multimodal data sources (eg, adding imaging and genomics' information, as well as structured clinical data and unstructured texts), FMs might assist in providing more comprehensive insights to a wider set of healthcare problems.[58,59] For example, FMs could accurately forecast COVID-19 cases and hospitalizations using real-time, complex, nonnumerical information, such as textual policies and genomic surveillance data, previously unattainable in traditional forecasting models.[60] Although challenges such as accuracy and reliability persist, advancements in fine-tuning, specialized models, and multimodal approaches suggest a path toward more comprehensive and accurate data sets. This work is needed to improve the comprehensiveness of value assessments for healthcare interventions.

Applications of Generative AI to Dossier Development and Reporting

Generative AI and FMs can be used to enhance the efficiency of dossier development for pharmaceutical product reporting and submissions to HTA agencies.[15] FMs excel at writing, can follow instructions for required styles, and can mimic the same style provided in other documents.[61,62] By automating the collation and presentation of evidence, generative AI might reduce the time and resources needed to produce comprehensive reports. Additionally, because FMs are language agnostic, they can generate documents suitable for different countries, facilitating international submissions and communications. FMs can also tailor communication materials to specific stakeholders, creating customized messages and visual aids that effectively convey the results of HEOR studies.[63] This makes them useful for example in developing lay summaries of technical reports.[15] Users must consider several limitations for effective FM implementation in dossier development. A primary concern is accuracy because FMs can generate plausible but incorrect information (eg, hallucinations), particularly with complex or nuanced scientific data. These inaccuracies risk undermining the credibility of submissions to regulatory agencies or HTA bodies, emphasizing the need for rigorous review processes to mitigate such risks.

Evaluating Research Output Using Generative AI

As these examples illustrate, generative AI is being applied across various HEOR domains; yet, assessing the quality and reliability of its outputs remains a critical challenge. Although generative AI holds significant promise for HEOR, most applications remain in early developmental stages. Of these, SLRs have progressed the furthest beyond proof of concept, whereas others require further validation to ensure reliability and practical utility. At this stage, it is difficult to separate limitations stemming from user expertise from those intrinsic to generative AI tools. Output quality often depends on user interaction, particularly in tasks such as SLRs or economic modeling. For instance, better prompting can enhance results, making user expertise a critical factor.[44] Therefore, it might be more practical to assess output quality using independent metrics that evaluate overall performance rather than separating user influence from the model's capabilities. To support this, evaluation frameworks are under development and will likely be helpful for both authors and reviewers.[47,48]

Table 2. Approaches to improve the quality of generative AI outputs in HEOR.

Technique/ approach	Description	Examples/ applications
Prompt engineering	Methods to refine input prompts for generative AI models to enhance output accuracy and comprehensiveness.	Examples include zero-shot, one-shot learning and chain-of-thought prompting to solve multistep problems or more complex scenarios.
Retrieval-augmented generation (RAG)	Models augment their generative process by retrieving external, domain-specific knowledge to improve accuracy and factuality.	Integration of any data sources (eg, publications, reports, and images) for answering specific queries; Bing or ChatGPT querying external websites for fact-checking and validation.
Model fine-tuning	Adjusting pretrained models on specific data sets for enhanced performance in specialized tasks to improve accuracy and completeness.	Fine-tuning on proprietary datasets for custom task performance (eg, domain-specific text, such as clinical EHRs); use of reinforcement learning with human feedback (RLHF) to improve factual accuracy.
Domain-specific FMs	Development and application of models trained on domain-specific corpora for increased accuracy and comprehensiveness.	Examples include GatorTron or BioClinicalBERT for specialized medical language understanding.
Multiple agents	Deploying distinct agents specialized in different tasks to work collaboratively. This will increase productivity and speed.	An example would be a retrieval agent for sourcing data or references, a summarization agent to condense information, and an analysis agent for data-driven insights, each working in tandem to solve complex problems.

AI indicates artificial intelligence; EHR, electronic health record; HEOR, health economics and outcomes research.

Techniques to improve the use of generative AI in HEOR

This section examines several more advanced techniques to enhance generative AI performance in HEOR. Strategies such as prompt engineering and retrieval-augmented generation (RAG) improve accuracy, factuality, and comprehensiveness. Model fine-tuning and domain-specific FMs ensure contextually relevant outputs for specialized tasks. These methods, applicable across use

Box 2
Prompting examples for systematic review of treatments for hepatitis C virus.

Zero-shot prompt:
"Screen the following abstracts and classify them as relevant or irrelevant to the systematic review of treatments for Hepatitis C Virus (HCV), based on study type, population, and outcomes. Provide a binary response for each abstract: relevant or irrelevant."
Few-shot prompt:
"Here are a few examples of abstracts that have already been classified. Use these examples to guide your classification of new abstracts for the HCV treatment review.
- Example 1: This randomized controlled trial evaluates the effectiveness of a new antiviral drug for HCV in patients with genotype 1. The study includes relevant clinical outcomes such as sustained virologic response (SVR) and adverse events. → Relevant
- Example 2: This study examines the prevalence of HCV in a specific geographic region without discussing treatment outcomes. → Irrelevant
- Now, based on these examples, classify the following abstracts as relevant or irrelevant to the systematic review."
Chain-of-thought prompt:
"Read the following abstract and explain your reasoning step by step before making a classification decision for the systematic review on HCV treatments. Focus on the type of study, the population studied, and whether the abstract includes treatment outcomes relevant to HCV care."
- Example Abstract: This observational study analyzes the long-term efficacy of a novel direct-acting antiviral (DAA) treatment in HCV patients across multiple genotypes. Outcomes include Sustained Virologic Response (SVR) and patient-reported quality of life measures.

The FM might provide an answer like this: "The study is observational, which can be included in a systematic review depending on the outcomes. The focus is on HCV treatment and includes important clinical outcomes like SVR and quality of life, which are directly relevant to the review question."
Persona pattern prompting:
"You are an experienced health economist tasked with identifying studies for a systematic review on HCV treatments. Your goal is to prioritize studies with robust methodologies that report clinical outcomes such as SVR and side effects. Classify the following abstract as relevant or irrelevant, considering your expertise in evaluating economic and clinical evidence."

cases, help address key challenges in accuracy and reliability. Table 2 summarizes key approaches to enhance the quality of generative AI outputs in HEOR.

Prompt engineering

Prompt engineering involves crafting specific inputs to optimize the quality of FM outputs.[64] It includes designing instructions, or prompts, that guide FMs to produce specific outputs through various strategies, such as zero-shot, few-shot, chain-of-thought, and persona pattern prompting.[65-67] Zero-shot prompting enables FMs to respond to novel queries based solely on the question posed without providing any examples (zero), whereas few-shot prompting enhances accuracy by offering a few relevant examples (few shot).[68] Chain-of-thought prompting facilitates complex decision making by guiding the model to display its reasoning process step by step.[69] Persona pattern prompting tailors outputs to align with the expertise and style expected of particular professional personas, such as health economists or policy makers.[64] Box 2 provides some examples of different types of prompts. Several prompt engineering techniques have been deployed in the conduct of SLRs and economic modeling.[27,34,35,44] However, prompt engineering is not without limitations. Zero-shot and few-shot prompting may miss important nuances in complex tasks, chain-of-thought can generate overly verbose outputs, and persona pattern prompting requires accurate replication of professional personas, which can be challenging.[64]

Model fine-tuning

Fine-tuning is a specialized technique in which a pretrained FM undergoes additional training using targeted data sets to refine its capabilities for specific tasks.[70] Fine-tuning might also take the shape of instruction tuning, using high-quality instruction-response pairs (ie, pairs of questions and answers, known to be correct).[71,72] Fine-tuning can significantly improve the model's performance on niche or complex tasks by adjusting its parameters to better reflect the unique needs of the application.[52] Self-improving feedback loops and reinforcement learning from human feedback is a form of fine-tuning.[71] Self-improving feedback loops iteratively refine prompts based on the outputs received, at the request of the model, to enhance model performance over time. Reinforcement learning from human feedback also has limitations which have been described in detail elsewhere.[73] Fine-tuning is different from training a specialized model from scratch. An example in HEOR is Bio-SIEVE, an FM designed to automate systematic reviews, specifically title and abstract screening. Instruction tuned on LLaMA and Guanaco models, Bio-SIEVE classifies studies for inclusion or exclusion based on predefined criteria and provides reasoning for exclusion, improving the efficiency of systematic reviews across medical domains.[29] Although fine-tuning offers significant benefits, it is not without challenges. It can be resource intensive, requiring substantial computational power and expert knowledge to perform the fine-tuning. There is also a risk of overfitting to the specific data used for training, which might limit the model's generalizability to other tasks.[74]

Domain-specific FMs

Domain-specific FMs are tailored for particular tasks or domains, such as biomedical research or healthcare, by utilizing domain-specific data sets to train the FM in more specific domains.[75] For example, in healthcare, these models can be trained on clinical notes from EHRs and other biomedical literature and texts, enhancing their performance on tasks such as clinical named entity recognition, reasoning tasks, question and answering, and text summarization. Domain-specific FMs can be trained from scratch using domain-specific data. For example, NYUTron, a large language model trained on unstructured clinical notes to predict important clinical outcomes, such as hospital mortality and length of stay, showed improvement compared with traditional models.[55] GatorTron and GatorTronGPT are a

generative clinical FM developed using GPT-3 architecture and trained on clinical text from clinical departments and patients at the University of Florida Health.[54,56] Both NYUTron and GatorTron have been shown to achieve superior performance in clinical NLP tasks, such as named entity recognition, compared with their general domain counterparts. In addition, domain-specific FMs can be built by continuously pretraining of open-domain FMs (eg, LLaMA) using domain-specific data. Examples of specialized models in this category include Me-LLaMA[57] and BioClinicalBERT, which are based on LLaMA and BERT, respectively, and are continuously trained on biomedical and clinical texts.[76] The enhanced performance of domain-specific FMs comes with limitations, including high training costs and technical demands, which may be prohibitive for smaller organizations. However, some models, such as GatorTron, are freely available to researchers under a license agreement.[54] Researchers should carefully balance the benefits of improved performance against the additional burdens of cost, learning time, and required expertise.

RAG

RAG is a sophisticated method that combines the broad knowledge base of FMs with precise, domain-specific data retrieval.[77] Conceptually a RAG system retrieves more up-to-date information, or task-specific information from external knowledge or data sources than the FM was pretrained on.[78] For example, a generative AI solution can use RAG to verify facts by accessing external databases or websites in real-time, ensuring that responses not only draw from a vast internal data set but are also cross verified with the latest external reference.[79] This capability significantly enhances the accuracy and reliability of the information provided, especially in rapidly evolving fields in which the FM might not have been trained on the most up-to-date data.[78,80] Unless carefully managed, a limitation of RAG could be its lack of ability in handling conflicting information retrieval, which can detrimentally affect RAG's output quality and is an active area of research.[77,80,81]

Agent-based techniques in generative AI for HEOR

Agents in generative AI are intelligent systems designed to perform specific tasks autonomously and interactively by combining a LLM or other FMs with additional capabilities, such as memory, task execution, interaction with external tools or data sources, and even collaboration with other agents.[82,83] Unlike standalone LLMs, which often handle individual language tasks (eg, text summarization), agents coordinate multiple components to execute complex workflows. For instance, virtual assistants, such as Alexa or Siri, powered with generative AI can now act as agents by responding to commands and performing tasks such as answering questions using integrated tools and external data sources.[84] Open-source tools are available for connecting LLMs to external resources.[85] They enable the creation of task-specific agents capable of automating workflows, such as retrieving data, synthesizing information, or running analyses. For example, Microsoft's AutoGen,[86] is an open-source tool designed for building AI agents and enabling cooperation among multiple agents. These tools allow agents to perform HEOR tasks more efficiently by interacting with external systems and retaining task context.

In HEOR, AI agents have significant potential to automate time-intensive tasks. For instance, in evidence synthesis, agents can retrieve and summarize research findings, streamlining SLRs. In economic modeling, they can extract and organize cost and utility data from real-world evidence, accelerating data integration. Agents can also aid dossier development by automating data collation and report formatting, reducing manual effort and enhancing consistency. For example, Generative pretrained transformer (GPT) Researcher, an autonomous agent, can conduct comprehensive web research based on text prompts and produce detailed, factual reports with citations.[87] However, challenges remain, including the technical expertise needed for development, ensuring the transparency and traceability of outputs, and addressing risks such as errors or hallucinations. Although agents show promise for HEOR, they are still in early development and require further refinement to ensure reliability and scalability.[82,83]

Application Programming Interfaces (APIs) and Their Role in Generative AI for HEOR

APIs are software intermediaries that enable communication between applications, facilitating structured data exchange based on predefined protocols. APIs facilitate the integration of LLMs into HEOR tasks by providing automation and structured data retrieval and ensuring seamless interoperability with external databases and analytical tools.[88,89]

API-based workflows require structured inputs to prevent failures and ensure data consistency. Additionally, security risks, such as prompt injection attacks, necessitate robust validation and access controls, particularly when handling sensitive data.[90] Compared with chatbots, which offer limited customization and workflow integration, APIs provide a scalable, modular, and secure foundation for incorporating LLMs into HEOR workflows, making them the preferred approach for building meaningful toolchains.

Limitations of Generative AI as Applied to HEOR

Although generative AI and FMs present promising opportunities in HEOR, several limitations must be acknowledged to provide a balanced understanding about the appropriate use of these technologies.

Scientific rigor in generative AI: Accuracy and reproducibility

The validity and reliability of generative AI tools in scientific research depend on their accuracy and reproducibility. Accuracy ensures that outputs are factual, error-free, and comprehensive,[47] whereas reproducibility allows findings to be independently verified under consistent conditions, building confidence in AI-generated insights.[91,92] Both are crucial for maintaining the credibility and utility of these tools in HEOR. Instances of inaccuracies in FM outputs, such as hallucinations—factually incorrect or fabricated information—have been documented, including examples such as non-existent citations generated during literature searches.[24,40] In the RWE space, one study reported less than 50% accuracy when mapping unstructured text to correct codes.[53] Similarly, Chhatwal et al found significant variability in how FMs represented disease progression in a Markov model for hepatitis C.[42] To improve accuracy, strategies such as refining prompts and using RAG have been effective in reducing hallucinations and increasing contextual accuracy.[81,93] Domain-specific FMs, such as GatorTron and Me-LLaMA, trained on large clinical data sets, and fine-tuning with domain-specific data have also enhanced performance in specialized research tasks.[56,57,59–61] Reproducibility, a cornerstone of scientific research, remains a challenge because of the black-box nature of FMs. AI models learn patterns from vast data sets, introducing variability in outputs based on input changes. Ensuring reproducibility will require an open sharing of data, code, and results, as well as adopting standardized reporting and transparency practices.[91,92] However, the probabilistic nature of these models means some variability will persist. HEOR-focused evaluation frameworks and reporting standards offer promising

paths to improving both accuracy and reproducibility, providing structured guidance for the use of generative AI in HEOR.[47,48]

Bias and fairness in generative AI

Bias and fairness are important considerations in the development and application of generative AI tools, including in HEOR.[3] Bias refers to systematic differences introduced during model development or deployment that can perpetuate inequities, whereas fairness ensures equitable operation across diverse populations and contexts.[94-97] FMs can propagate or amplify biases arising from stages such as training data selection or deployment, potentially causing harm to individuals and communities.[95,96,98] An interesting example highlighting concerns associated with bias comes from a study by Giyocha et al,[99] in which an AI model accurately predicted a patient's self-reported race from medical imaging across various modalities—an ability beyond clinical experts.[3,99] However, the study could not determine the specific features the model used, raising ethical questions about fairness. For instance, if models implicitly use race as a factor, they risk biased treatment recommendations, potentially exacerbating health disparities. Several strategies have been proposed to manage the risk of bias in FM outputs. Published surveys explore methodologies to evaluate fairness, identify sources of bias, and implement corrective actions.[94,95,97,100,101] These approaches include fairness metrics, such as demographic parity and equalized odds, bias audits, adversarial debiasing, reweighting training data, and fine-tuning with domain-specific data sets.[94,95,97] Synthetic data sets have also shown promise in improving fairness.[102] Research on evaluating and mitigating bias in FMs applied to HEOR are an active area of research.3,97

Technical and operational considerations in generative AI for HEOR

The integration of FMs into HEOR is not without technical and operational challenges, including compliance with national regulations, ensuring security and privacy, balancing open-source and proprietary models, addressing deployment issues, managing costs, and integrating FMs into existing workflows.

First, generative AI tools must comply with national and local regulations, such as the EU AI Act, Health Insurance Portability and Accountability Act in the United States, and General Data Protection Regulation in the European Union.[103] FMs can pose risks by memorizing and reproducing sensitive data, including Protected Health Information.[104,105] Safeguards such as encryption, anonymization, access controls, and compliance with standards such as the Federal Information Security Management Act and the HITRUST Framework are critical to mitigate these risks.[106,107] Federated analytics offers a promising solution by enabling analysis across multiple sources without centralizing data, preserving privacy while ensuring analytical rigor.[108]

Second, the choice between open-source and proprietary FMs plays a crucial role in shaping their adoption and use within HEOR.[9] Open-source models, such as GPT-Neo, LLaMA, and DeepSeek, offer transparency and customization but require substantial resources, expertise, and maintenance to keep pace with the field. Proprietary models, such as GPT-4, are easier to deploy and are optimized for performance but come with limitations such as vendor lock-in, privacy concerns, and limited transparency, complicating bias and fairness assessments.[109]

Third, deployment strategies significantly affect FM adoption. Cloud-based solutions offer scalability and accessibility but require stringent security measures to ensure compliance with regulations such as Health Insurance Portability and Accountability Act and General Data Protection Regulation because of potential data breach risks.[110,111] In practice, many organizations opt for self-managed cloud environments—such as AWS or Azure infrastructure as a service— rather than relying on vendor-managed software as a service or platform as a service cloud solutions, allowing for greater control over security and compliance. Although fully on-premises deployment on dedicated infrastructure is uncommon, some organizations may still pursue this approach for specific regulatory or data sensitivity reasons. However, both self-managed cloud and on-premises approaches come with cost challenges, including licensing fees, infrastructure expenses, and the need for specialized expertise. Organizations must assess their security, compliance, and cost considerations to determine the most suitable deployment strategy.

Finally, integrating FMs into HEOR workflows involves both technical and operational challenges. Platforms and tools such as LangChain or API-based systems offer the building blocks to automate tasks such as systematic reviews or economic modeling but require advanced expertise and are often hindered by the current absence of standardized APIs and interoperability protocols.[85] Simplifying deployment tools and creating user-friendly interfaces or mature products tailored to specific HEOR tasks will be important for wider adoption. Organizational barriers, including resistance to change, reliance on traditional methods, limited AI expertise, and high implementation costs, further hinder adoption. Providing training and upskilling opportunities for HEOR professionals will be essential to overcoming these challenges and enabling the effective integration of FMs into HEOR workflows.[112]

Conclusions

This article introduces key concepts in generative AI for the HEOR community, providing a foundation for understanding and engaging with this transformative technology. Generative AI and FMs offer significant potential to enhance HEOR by streamlining workflows, automating complex tasks, and introducing innovative solutions. However, this promise comes with challenges, including ensuring scientific accuracy and reproducibility, addressing bias and fairness, and managing technical and operational complexities. Advanced techniques, such as prompt engineering and RAG, may mitigate some of these limitations, but the field is still in its early stages of optimizing these tools. For HEOR professionals, the path forward requires embracing generative AI with a commitment to rigorous validation, interdisciplinary collaboration, and ongoing learning. Thoughtfully integrating these tools can support equitable and impactful healthcare decisions, ultimately improving patient outcomes.

Glossary

- Agents: Autonomous or semiautonomous systems that use LLMs to perform tasks with minimal human input, processing data, generating text, adapting to new information, and refining outputs based on objectives.
- Artificial intelligence (AI): a broad field of computer science that aims to create intelligent machines capable of performing tasks typically requiring human intelligence.
- Deep learning: a subset of machine learning algorithms that uses multilayered neural networks, called deep neural networks. Those algorithms are the core behind the majority of advanced AI models.
- Foundation model: large-scale pretrained models that serve a variety of purposes. These models are trained on broad data at scale and can adapt to a wide range of tasks and domains with further fine-tuning.

- Generative AI: AI systems capable of generating text, images, or other content based on input data, often creating new and original outputs.
- Generative pretrained transformer (GPT): a specific series of FMs created by OpenAI based on the Transformer architecture, which is particularly well suited for generating human-like text.
- Large language model: a specific type of FM trained on massive text data that can recognize, summarize, translate, predict, and generate text and other content based on knowledge gained from massive data sets.
- Machine learning (ML): a field of study within AI that focuses on developing algorithms that can learn from data without being explicitly programmed.
- Multimodal AI: an AI model that simultaneously integrates diverse data formats provided as training and prompt inputs, including images, text, bio-signals, -omics data and more.
- Prompt: the input given to an AI system, consisting of text or parameters that guide the AI to generate text, images, or other outputs in response.
- Prompt engineering: creating and adapting prompts (input) to instruct AI models to generate specific output.
- Supervised learning: a machine learning approach in which models are trained on labeled data, pairing inputs with known outputs. This enables the model to learn patterns for tasks such as predicting healthcare costs, diagnosing diseases, or classifying images.
- Token: a unit of text processed by an AI model, which can be a word, subword, or character. AI models convert input text into tokens to generate or interpret language. The number of tokens affects model performance, cost, and the ability to process longer texts efficiently.
- Unsupervised learning: A machine learning approach in which models are trained on unlabeled data to uncover hidden patterns or structures, such as clustering patients with similar health outcomes.

Author Disclosures

Author disclosure forms can be accessed below in the Supplemental Material section. The views expressed are those of the authors and not those of their employing or funding organizations. Dr Fleurence contributed to this article in her personal capacity. The views expressed are her own and do not necessarily represent the views of the National Institutes of Health or the United States Government. Drs Fleurence, Dawoud, and Chhatwal are editors for *Value in Health* and had no role in the peer-review process of this article.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2025.04.2167.

Article and Author Information

Accepted for Publication: April 25, 2025

Published Online: June 16, 2025

doi: https://doi.org/10.1016/j.jval.2025.04.2167

Author Affiliations: National Institute of Biomedical Imaging and Bioengineering, Bethesda, MD, USA (Fleurence); Value Analytics Labs, Boston, MA, USA (Fleurence); Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, USA (Wang); Intelligent Medical Objects, Rosemont, IL, USA (Wang); Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, FL, USA (Bian); Biomedical Informatics, Clinical and Translational Science Institute, University of Florida, FL, USA (Bian); Office of Data Science and Research Implementation, University of Florida Health, Gainesville, FL, USA (Bian); Regenstrief Institute, Indianapolis, IN, USA (Bian); Biostatistics and Health Data Science, School of Medicine, Indiana University, Indianapolis, IN, USA (Bian); ISPOR, The Professional Society for Health Economics and Outcomes Research, Lawrenceville, NJ, USA (Higashi); Center for Health and Humanitarian Systems, Georgia Institute of Technology, Atlanta, GA, USA (Ayer); Value Analytics Labs, Boston, MA, USA (Ayer); Institute Department of Biomedical Informatics and Data Science, School of Medicine, Yale University, New Haven, CT, USA (Xu); National Institute for Health and Care Excellence, London, England, UK (Dawoud); Cairo University, Faculty of Pharmacy, Cairo, Egypt (Dawoud); Institute for Technology Assessment, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA (Chhatwal); Center for Health Decision Science, Harvard University, Boston, MA, USA (Chhatwal).

Correspondence: Rachael L. Fleurence, PhD, Value Analytics Labs, 100 Cambridge St Suite 1400, Boston, MA 02114, USA. Email: Rachael.fleurence@nih.gov

Authorship Confirmation: All authors certify that they meet the ICMJE criteria for authorship.

Funding/Support: Dr Dawoud was partly supported by funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 82516 (Next Generation Health Technology Assessment [HTx] project). The other authors received no financial support for this research.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Acknowledgment: This manuscript was developed as part of the ISPOR Working Group on Generative AI. The authors thank the ISPOR Science office for their support and Sahar Alam for her excellent program management throughout the project.

REFERENCES

1. AIs will make health care safer and better. The Economist. https://www.economist.com/technology-quarterly/2024/03/27/ais-will-make-health-care-safer-and-better. Accessed June 5, 2025.
2. Telenti A, Auli M, Hie BL, Maher C, Saria S, Ioannidis JPA. Large language models for science and medicine. *Eur J Clin Invest*. 2024;54:e14183.
3. Fleurence RL, Bian J, Wang X, et al. Generative artificial intelligence for health technology assessment: opportunities, challenges, and policy considerations: an ISPOR working group report. *Value Health*. 2024;28:175–183.
4. Howell MD, Corrado GS, DeSalvo KB. Three epochs of artificial intelligence in health care. *JAMA*. 2024;331(3):242–244.
5. Hinton GE, Osindero S, Teh YW. A fast learning algorithm for deep belief nets. *Neural Comput*. 2006;18(7):1527–1554.
6. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583–589.
7. Remarkable progress has been made in understanding the folding of proteins. The Economist. https://www.economist.com/leaders/2021/07/31/remarkable-progress-has-been-made-in-understanding-the-folding-of-proteins. Accessed June 5, 2025.
8. The Nobel Prize in Chemistry. The Nobel Prize. https://www.nobelprize.org/prizes/chemistry/. Accessed October 9, 2024.
9. Schneider J, Meske C, Kuss P. Foundation models a new paradigm for artificial intelligence. *Bus Inf Syst Eng*. 2024;66(2):221–231.
10. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. Accessed June 5, 2025.
11. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877–1901.
12. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. *Nat Med*. 2023;29(8):1930–1940.
13. Introducing ChatGPT. Open AI. https://openai.com/blog/chatgpt. Accessed June 5, 2025.
14. Zhao WX, Zhou K, Li J, et al. A survey of large language models. *arXiv*. https://doi.org/10.48550/arXiv.2303.18223
15. Use of AI in evidence generation: NICE position statement. National Institute for Health and Care Excellence. https://www.nice.org.uk/about/what-we-do/

our-research-work/use-of-ai-in-evidence-generation–nice-position-statement. Accessed September 20, 2024.
16. NICE statement of intent for artificial intelligence (AI). National Institute for Health and Care Excellence. https://www.nice.org.uk/corporate/ecd12/resources/nice-statement-of-intent-for-artificial-intelligence-ai-pdf-40464270623941. Accessed December 16, 2024.
17. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science*. 2015;349(6245):255–260.
18. Breiman L. Random Forests. *Mach Learn*. 2001;45(1):5–32.
19. Vapnik V. *The Nature of Statistical Learning Theory*. Berlin, Germany: Springer-Verlag; 1995.
20. Padula WV, Kreif N, Vanness DJ, et al. Machine learning methods in health economics and outcomes research-the PALISADE checklist: a good practices report of an ISPOR task force. *Value Health*. 2022;25(7):1063–1080.
21. Hu Y, Chen Q, Du J, et al. Improving large language models for clinical named entity recognition via prompt engineering. *J Am Med Inform Assoc*. 2024;31(9):1812–1820.
22. Tsertsvadze A, Chen YF, Moher D, Sutcliffe P, McCarthy N. How to conduct systematic reviews more expeditiously? *Syst Rev*. 2015;4:160.
23. Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd ed. Chichester, United Kingdom: John Wiley & Sons; 2019.
24. Qureshi R, Shaughnessy D, Gill KAR, Robinson KA, Li T, Agai E. Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Syst Rev*. 2023;12(1):72.
25. Khraisha Q, Put S, Kappenberg J, Warraitch A, Hadfield K. 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. *Res Synth Methods*. 2024;15(4):616–626.
26. Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. Automated paper screening for clinical reviews using large language models: data analysis study. *J Med Internet Res*. 2024;26:e48996.
27. Tran VT, Gartlehner G, Yaacoub S, et al. Sensitivity and specificity of using GPT-3.5 turbo models for title and abstract screening in systematic reviews and meta-analyses. *Ann Intern Med*. 2024;177(6):791–799.
28. Landschaft A, Antweiler D, Mackay S, et al. Implementation and evaluation of an additional GPT-4-based reviewer in PRISMA-based medical systematic literature reviews. *Int J Med Inform*. 2024;189:105531.
29. Robinson A, Thorne W, Wu BP, et al. Bio-sieve: exploring instruction tuning large language models for systematic review automation. *arXiv*. https://doi.org/10.48550/arXiv.2308.06610
30. Hasan B, Saadi S, Rajjoub NS, et al. Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment. *BMJ Evid Based Med*. 2024;29(6):394–398.
31. Lai H, Ge L, Sun M, et al. Assessing the risk of bias in randomized clinical trials with large language models. *JAMA Netw Open*. 2024;7(5):e2412687.
32. Lee K, Paek H, Huang LC, et al. SEETrials: leveraging large language models for safety and efficacy extraction in oncology clinical trials. *Inform Med Unlocked*. 2024;50:101589.
33. Schopow N, Osterhoff G, Baur D. Applications of the natural language processing tool ChatGPT in clinical practice: comparative study and augmented systematic review. *JMIR Med Inform*. 2023;11:e48933.
34. Gartlehner G, Kahwati L, Hilscher R, et al. Data extraction for evidence synthesis using a large language model: a proof-of-concept study. *Res Synth Methods*. 2024;15(4):576–589.
35. Reason T, Benbow E, Langham J, Gimblett A, Klijn SL, Malcolm B. Artificial intelligence to automate network meta-analyses: four case studies to evaluate the potential application of large language models. *Pharmacoecon Open*. 2024;8(2):205–220.
36. Jansen JP, Fleurence R, Devine B, et al. Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1. *Value Health*. 2011;14(4):417–428.
37. Yun HS, Pogrebitskiy D, Marshall IJ, Wallace BC. Automatically extracting numerical results from randomized controlled trials with large language models. *arXiv*. https://doi.org/10.48550/arXiv.2405.01686.
38. Huang Y, Wu R, He J, Xiang Y. Evaluating ChatGPT-4.0's data analytic proficiency in epidemiological studies: a comparative analysis with SAS, SPSS, and R. *J Glob Health*. 2024;14:04070.
39. Wu Y, Klijn S, Teitsson S, Malcolm B, Jones C, Rawlinson W. Innovations in automated survival curve selection and reporting of survival analyses through generative AI. https://www.ispor.org/heor-resources/presentations-database/presentation-paper/euro2024-4003/19093/innovations-in-automated-survival-curve-selection-and-reporting-of-survival-analyses-through-generative-ai. Accessed January 19, 2025.
40. Jin Q, Leaman R, Lu Z. Retrieve, summarize, and verify: how will ChatGPT affect information seeking from the medical literature? *J Am Soc Nephrol*. 2023;34(8):1302–1304.
41. Smela B, Łukiewicz B, Gawlik K, Clay E, Boyer L, Toumi M. Balancing feasibility, time, and comprehensiveness: approaches to rapid reviews of health economic models. https://www.ispor.org/heor-resources/presentations-database/presentation/intl2024-3896/138795. Accessed June 5, 2025.
42. Chhatwal J, Yildrim IF, Balta D, et al. Can large language models generate conceptual health economic models?. https://www.ispor.org/heor-resources/presentations-database/presentation/intl2024-3898/139128. Accessed June 5, 2025.
43. Chhatwal J, Yildirim IF, Samur S, Bayraktar E, Ermis T, Ayer T. Development of de novo health economic models using generative AI. *Value Health*. 2024;27:S7.
44. Reason T, Rawlinson W, Langham J, Gimblett A, Malcolm B, Klijn S. Artificial intelligence to automate health economic modelling: a case study to evaluate the potential application of large language models. *Pharmacoecon Open*. 2024;8(2):191–203.
45. Chhatwal J, Samur S, Yildirim IF, Bayraktar E, Ermis T, Ayer T. Fully replicating published Markov health economic models using generative AI. *Value Health*. 2024;27:S102.
46. Briggs AH, Weinstein MC, Fenwick EA, et al. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force–6. *Value Health*. 2012;15(6):835–842.
47. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*. 2025;333(4):319–328.
48. Fleurence RL, Dawoud D, Bian J, et al. ELEVATE-GenAI: Reporting Guidelines for the use of Large Language Models in Health Economics and Outcomes Research: an ISPOR Working Group on Generative AI Report. *arXiv*. https://doi.org/10.48550/arXiv.2501.12394
49. Fleurence R, Kent S, Adamson B, et al. Assessing real-world data from electronic health records for health technology assessment – the SUITABILITY checklist: a good practices report from an ISPOR task force. *Value Health*. 2024;27(6):692–701.
50. Lee K, Liu Z, Chandran U, et al. Detecting ground glass opacity features in patients with lung cancer: automated extraction and longitudinal analysis via deep learning-based natural language processing. *JMIR AI*. 2023;2:e44537.
51. Guo LL, Fries J, Steinberg E, et al. A multi-center study on the adaptability of a shared foundation model for electronic health records. *npj Digit Med*. 2024;7(1):171.
52. Cohen AB, Waskom M, Adamson B, Kelly J, GA. Using large language models to extract PD-L1 testing details from electronic health records. https://www.ispor.org/heor-resources/presentations-database/presentation/intl2024-3898/136019. Accessed June 5, 2025.
53. Soroush A, Glicksberg BS, Zimlichman E, et al. Large language models are poor medical coders — benchmarking of medical code querying. *NEJM AI*. 2024;1(5):AIdbp2300040.
54. Peng C, Yang X, Chen A, et al. A study of generative large language model for medical research and healthcare. *npj Digit Med*. 2023;6(1):210.
55. Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. *Nature*. 2023;619(7969):357–362.
56. Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med*. 2022;5(1):194.
57. Xie Q, Chen Q, Chen A, et al. Me-LLaMA: foundation large language models for medical applications. *Res Sq*. https://doi.org/10.21203/rs.3.rs-4240043/v1
58. Yang L, Xu S, Sellergren A, et al. Advancing multimodal medical capabilities of Gemini. *arXiv*. https://doi.org/10.48550/arXiv.2405.03162
59. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med*. 2022;28(1):31–38.
60. Du H, Zhao J, Zhao Y, et al. Advancing real-time pandemic forecasting using large language models: a COVID-19 case study. *arXiv*. https://doi.org/10.48550/arXiv.2404.06962
61. Smith GR, Bello C, Bialic-Murphy L, et al. Ten simple rules for using large language models in science, version 1.0. *PLoS Comput Biol*. 2024;20(1):e1011767.
62. Team Gemini, Anil R, Borgeaud S, et al. Gemini: a family of highly capable multimodal models. *arXiv*. https://doi.org/10.48550/arXiv.2312.11805
63. August T, Lo K, Smith NA, Reinecke K. Know your audience: the benefits and pitfalls of generating plain language summaries beyond the "general" audience. https://dl.acm.org/doi/full/10.1145/3613904.3642289. Accessed June 5, 2025.
64. Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Med Inform*. 2024;12:e55318.
65. Schulhoff S, Ilie M, Balepur N, et al. The prompt report: a systematic survey of prompting techniques. *arXiv*. https://doi.org/10.48550/arXiv.2406.06608
66. Lin Z. How to write effective prompts for large language models. *Nat Hum Behav*. 2024;8(4):611–615.
67. Lin Z. Why and how to embrace AI such as ChatGPT in your academic life. *R Soc Open Sci*. 2023;10(8):230658.
68. Kojima T, Shane Gu S, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. https://dl.acm.org/doi/10.5555/3600270.3601883. Accessed June 5, 2025.
69. Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv Neural Inf Process Syst*. 2022;35:24824–24837.
70. Chung HW, Hou L, Longpre S, et al. Scaling instruction-finetuned language models. *J Mach Learn Res*. 2024;25(70):1–53.
71. Ouyang L. Training language models to follow instructions with human feedback. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html. Accessed June 5, 2025.

72. Yue X, Zheng T, Zhang G, Chen W. Mammoth2: scaling instructions from the web. *arXiv*. https://doi.org/10.48550/arXiv.2405.03548
73. Casper S, Davies X, Shi C, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv*. https://doi.org/10.48550/arXiv.2307.15217
74. Szép M, Rueckert D, von Eisenhart-Rothe R, Hinterwimmer F. A practical guide to fine-tuning language models with limited data. *arXiv*. https://doi.org/10.48550/arXiv.2411.09539
75. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616(7956):259–265.
76. Alsentzer E, Murphy JR, Boag W, et al. Publicly available clinical BERT embeddings. *arXiv*. https://doi.org/10.48550/arXiv.1904.03323
77. Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: a survey. *arXiv*. https://doi.org/10.48550/arXiv.2312.10997
78. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html. Accessed June 5, 2025.
79. Retrieval augmented generation (RAG) and semantic search for GPTs. Open AI. https://help.openai.com/en/articles/8868588-retrieval-augmented-generation-rag-and-semantic-search-for-gpts. Accessed January 20, 2025.
80. Lin XV, Chen X, Chen M, et al. Ra-dit: retrieval-augmented dual instruction tuning. *arXiv*. https://doi.org/10.48550/arXiv.2310.01352
81. Yu H, Gan A, Zhang K, Tong S, Liu Q, Liu Z. Evaluation of retrieval-augmented generation: a survey. *arXiv*. https://doi.org/10.48550/arXiv.2405.07437
82. Xi Z, Chen W, Guo X, et al. The rise and potential of large language model based agents: a survey. *arXiv*. https://doi.org/10.48550/arXiv.2309.07864
83. Cheng Y, Zhang C, Zhang Z, et al. Exploring large language model based intelligent agents: definitions, methods, and prospects. *arXiv*. https://doi.org/10.48550/arXiv.2401.03428
84. Clarke C, Krishnamurthy K, Talamonti W, Kang Y, Tang L, Mars J. One agent too many: user perspectives on approaches to multi-agent conversational AI. *arXiv*. https://doi.org/10.48550/arXiv.2401.07123
85. LangChain. https://www.langchain.com/. Accessed January 20, 2025.
86. AutoGen: open-source programming framework for agentic AI. Microsoft. https://www.microsoft.com/en-us/research/project/autogen/. Accessed January 20, 2025.
87. GPT Researcher. https://gptr.dev/. Accessed January 20, 2025.
88. How do I start exploring the OpenAI API. Open AI. https://help.openai.com/en/articles/4936851-how-do-i-start-exploring-the-openai-api. Accessed October 1, 2024.
89. Kumar D. Function calling in LLM. https://medium.com/@danushidk507/function-calling-in-llm-e537b286a4fd. Accessed March 14, 2025.
90. Wu Z, Gao H, He J, Wang P. The dark side of function calling: pathways to jailbreaking large language models. *arXiv*. https://doi.org/10.48550/arXiv.2407.17915
91. Beam AL, Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. *JAMA*. 2020;323(4):305–306.
92. Kapoor S, Cantrell EM, Peng K, et al. REFORMS: consensus-based recommendations for machine-learning-based science. *Sci Adv*. 2024;10(18):eadk3452.
93. Wei CH, Allot A, Lai PT, et al. PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Res*. 2024;52(W1):W540–W546.
94. Caton S, Haas C. Fairness in machine learning: a survey. *ACM Comput Surv*. 2024;56(7):1–38.
95. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv (CSUR)*. 2021;54(6):1–35.
96. Drukker K, Chen W, Gichoya J, et al. Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. *J Med Imaging (Bellingham)*. 2023;10(6):061104.
97. Yang Y, Lin M, Zhao H, Peng Y, Huang F, Lu Z. A survey of recent methods for addressing AI fairness and bias in biomedicine. *J Biomed Inform*. 2024;154:104646.
98. Gervasi S, Chen I, Smith-McLallen A, et al. The potential for bias in machine learning and opportunities for health insurers to address It. *Health Aff*. 2022;41(2):212–218.
99. Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health*. 2022;4(6):e406–e414.
100. Xu J, Xiao Y, Wang WH, et al. Algorithmic fairness in computational medicine. *EBiomedicine*. 2022;84:104250.
101. Huang Y, Guo J, Chen WH, et al. A scoping review of fair machine learning techniques when using real-world data. *J Biomed Inform*. 2024;151:104622.
102. Mosquera L, El Emam K, Ding L, et al. A method for generating synthetic longitudinal health data. *BMC Med Res Methodol*. 2023;23(1):67.
103. European Union. EU artificial intelligence act. https://artificialintelligenceact.eu/the-act/. Accessed May 22, 2024.
104. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule. *J Am Med Inform Assoc*. 2010;17(2):169–177.
105. Simon GE, Shortreed SM, Coley RY, et al. Assessing and minimizing re-identification risk in research data derived from health care records. *eGEMs (Wash DC)*. 2019;7(1):6.
106. HITRUST. https://hitrustalliance.net/hitrust-framework. Accessed January 20, 2025.
107. FISMA. https://csrc.nist.gov/topics/laws-and-regulations/laws/FISMA. Accessed January 20, 2025.
108. Ren C, Yu H, Peng H, et al. Advances and open challenges in federated foundation models. *arXiv*. https://doi.org/10.48550/arXiv.2404.15381
109. Lu S. Proprietary vs. open source foundation models. https://tolacapital.com/2023/05/15/foundationmodels?utm_source=chatgpt.com. Accessed January 20, 2025.
110. Luqman A, Mahesh R, Chattopadhyay A. Privacy and security implications of cloud-based ai services: a survey. *arXiv*. https://doi.org/10.48550/arXiv.2402.00896
111. Zandesh Z. Privacy, security, and legal issues in the health cloud: structured review for taxonomy development. *JMIR Form Res*. 2024;8:e38372.
112. Zemplenyi A, Tachkov K, Balkanyi L, et al. Recommendations to overcome barriers to the use of artificial intelligence-driven evidence in health technology assessment. *Front Public Health*. 2023;11:1088121.