

ISPOR Report

ELEVATE-GenAI: Reporting Guidelines for the Use of Large Language Models in Health Economics and Outcomes Research: An ISPOR Working Group Report

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

ABSTRACT

Objectives: Generative artificial intelligence (AI), particularly large language models (LLMs), holds significant promise for health economics and outcomes research (HEOR). However, standardized reporting guidance for LLM-assisted research is lacking. This article introduces the ELEVATE-GenAI framework and checklist—reporting guidelines specifically designed for HEOR studies involving LLMs.

Methods: The framework was developed through a targeted literature review of existing reporting guidelines, AI evaluation frameworks, and expert input from the ISPOR Working Group on Generative AI. It comprises 10 domains—including model characteristics, accuracy, reproducibility, and fairness and bias. The accompanying checklist translates the framework into actionable reporting items. To illustrate its use, the framework was applied to 2 published HEOR studies: one focused on a systematic literature review tasks and the other on economic modeling.

Results: The ELEVATE-GenAI framework offers a comprehensive structure for reporting LLM-assisted HEOR research, while the checklist facilitates practical implementation. Its application to the 2 case studies demonstrates its relevance and usability across different HEOR contexts.

Conclusions: Although the framework provides robust reporting guidance, further empirical testing is needed to assess its validity, completeness, usability, and generalizability across diverse HEOR use cases.

The ELEVATE-GenAI framework and checklist address a critical gap by offering structured guidance for transparent, accurate, and reproducible reporting of LLM-assisted HEOR research. Future work will focus on extensive testing and validation to support broader adoption and refinement.

Keywords: artificial intelligence, generative AI, large language model, reporting guidelines.

VALUE HEALTH. 2025; 28(11):1611–1625

Highlights

- This article addresses the lack of structured guidance for reporting research using large language models (LLMs) in health economics and outcomes research (HEOR) by introducing the ELEVATE-GenAI framework and checklist.
- The ELEVATE-GenAI framework and checklist provides a practical, domain-specific tool for systematically reporting the use of LLMs in HEOR research, emphasizing 10 domains, including transparency, accuracy, and reproducibility.
- The reporting guidelines promote rigorous reporting standards, enabling HEOR professionals to integrate LLMs responsibly, enhancing evidence synthesis, modeling, and real-world data generation in healthcare research.

Introduction

Artificial intelligence (AI) encompasses computational methods for tasks requiring human-like reasoning, learning, or decision making.¹ Natural language processing, a subfield of AI, enables machines to understand and generate human language.² Generative AI (Gen AI) models produce new content—such as text, code, or data—based on patterns in training data,^{3,4} with large language models (LLMs) emerging as especially impactful. Foundation models, such as Generative Pre-trained Transformer (GPT), Gemini, Claude, and LLaMA, trained on vast corpora via self-supervised learning, now support increasingly multimodal tasks across text, image, and other data modalities.^{5,6} The 2022 release of ChatGPT marked a major shift, expanding LLM access to broader user groups, including health economics and outcomes research (HEOR) researchers.^{3,7}

Gen AI, particularly LLMs, is rapidly transforming HEOR by augmenting traditionally labor-intensive tasks, such as systematic

reviews, model development, and evidence generation.^{3,8} However, the growing integration of LLMs into scientific workflows raises critical concerns around transparency, reproducibility, and trustworthiness—challenges for which HEOR-specific reporting standards do not yet exist.^{3,8}

In HEOR, LLMs are already being used to support systematic literature reviews (SLRs), health economic modeling (HEM), and real-world evidence (RWE) generation. These applications include tasks such as abstract screening, bias assessment, meta-analysis automation, parameter estimation, and transforming unstructured real-world data from electronic health records, imaging, and genomics into analyzable formats.^{9–31} Although these uses offer substantial promise, limitations such as hallucinations, data inaccuracies, and the need for human oversight underscore the importance of structured reporting practices.^{3,6,8}

Regulatory and health technology assessment (HTA) bodies have begun issuing guidance. For example, the US Food and Drug Administration (FDA) recently issued draft guidance proposing a

risk-based credibility assessment framework for AI in regulatory submissions, including LLMs³² and a perspective on the use of AI in its work.³³ The UK's National Institute for Health and Care Excellence (NICE) has also released both a Statement of Intent and a position statement outlining principles for generative AI use in HTA submissions,^{34,35} as has Canada's Drug Agency.³⁶

To address the lack of HEOR-specific reporting standards, the ISPOR Working Group on Gen AI developed the ELEVATE-GenAI framework. These provide structured criteria to help researchers transparently report how LLMs are used to generate or analyze evidence. Although applicable for evaluation, the primary aim is to support reproducible reporting and peer review. The guidelines target studies in which LLMs play a substantive role—such as in systematic reviews, economic modeling, or real-world data analysis—not those using AI for limited tasks such as editing or summarization. Researchers are encouraged to apply judgment based on the context of AI use.

The article begins by presenting the literature review that informed the framework's development. After a detailed overview of the framework and its domains, the guidelines are applied to 2 HEOR use cases—one in systematic review and one in economic modeling—to illustrate practical use. As a living guideline, ELEVATE-GenAI could evolve with community input and advances in generative AI. Future updates would be versioned and publicly available, with structured piloting and validation led by the ISPOR Working Group on Generative AI to ensure continued relevance, completeness and usability.

Methods

The ELEVATE-GenAI reporting guidelines were developed through a multistep process involving a targeted literature review, iterative framework construction, and initial application to published HEOR use cases.

Targeted Literature Review

A targeted literature review was conducted to identify existing evaluation frameworks, reporting guidelines, and governance principles relevant to the use of LLMs in healthcare and health research. Searches were conducted in PubMed (through January 31, 2025) and ArXiv (through December 31, 2024), and additional reporting guidelines were retrieved from the EQUATOR Network,³⁷ a clearing house for reporting guidelines. The search strategy, eligibility criteria, and PRISMA flow diagram are available in the [Supplemental Materials](https://doi.org/10.1016/j.jval.2025.06.018) found at <https://doi.org/10.1016/j.jval.2025.06.018>. Title and abstract screening were conducted by a single reviewer (R.F.) using predefined eligibility criteria. Full-text screening was conducted by R.F., with input from a second reviewer (J.C.) for uncertain cases. Data extraction was completed using a structured template to capture article title, purpose, and proposed reporting elements. Extraction was independently reviewed on a sub-sample of articles by additional co-authors (J.B., J.C., X.W.).

Framework Development

Findings from the targeted literature review informed the identification of key reporting domains for LLM use in HEOR. These were refined through iterative discussions within the ISPOR Working Group on Generative AI, drawing on technical literature, regulatory guidance, and real-world use cases. The framework was designed for flexibility across core HEOR applications—SLRs, HEM, and RWE—covering both high-level tasks and subtasks (eg, abstract screening, model specification). To test usability and relevance, the framework was applied to 2 published HEOR studies:

one focused on systematic review¹⁶ and one on economic modeling,²³ to assess domain coverage across different use cases.

The ELEVATE-GenAI framework is intended as a living guideline that will be refined through structured validation. Planned next steps include stakeholder consultation with researchers, industry experts, and regulatory bodies, piloting in active HEOR studies, and a formal Delphi process to assess the clarity, relevance, and utility of each reporting domain. These activities, modeled on best practices from prior guideline development efforts (eg, Preferred Reporting Items for Systematic Reviews and Meta-Analyses - Artificial Intelligence [PRISMA-AI]³⁸), will support broader adoption and ensure the framework remains scientifically rigorous, usable, and adaptable as the field of generative AI evolves.

Results

Literature Search Results

A total of 522 records were identified through PubMed and ArXiv searches. After title and abstract screening, 490 records were excluded, and 32 full-text articles were assessed for eligibility. Of these, 17 were excluded, resulting in 15 studies included in the final synthesis.^{3,4,39-51} An additional 6 reporting guidelines^{38,52-56} and 9 position statements or frameworks^{32,34,57-63} from international organizations, regulatory agencies, or HTA bodies (eg, NICE and FDA) were included yielding a total of 30 sources included in the literature review. The [Appendix 1](https://doi.org/10.1016/j.jval.2025.06.018) and [Appendix 2](https://doi.org/10.1016/j.jval.2025.06.018) in [Supplemental Materials](https://doi.org/10.1016/j.jval.2025.06.018) found at <https://doi.org/10.1016/j.jval.2025.06.018> provides the search strategy, eligibility criteria, PRISMA flow diagram, and a table summarizing the included studies and reports.

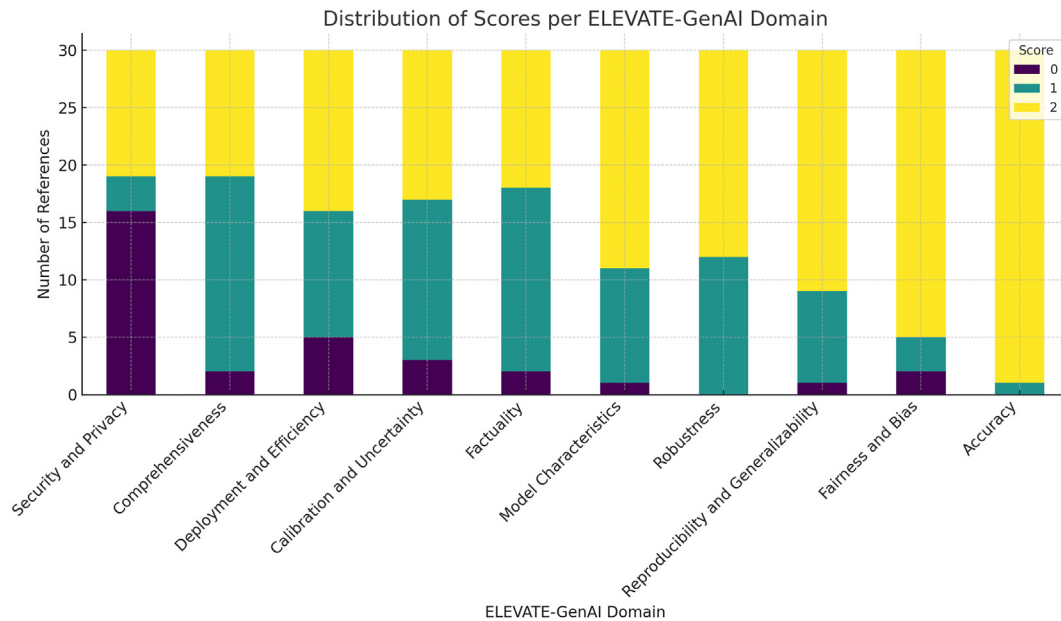
Overview of Literature Identified

The 15 studies proposing evaluation frameworks included systematic reviews, conceptual models, and benchmarking protocols across domains such as clinical research, general medicine, evidence synthesis, and health technology assessment.^{3,4,39-51} Nine guidance documents from national agencies, international organizations, and HTA bodies were identified.^{32,34,57-63} Although some focused broadly on AI/machine learning (ML) rather than on LLMs specifically, they were included for their relevance to responsible AI use in healthcare. Six reporting guidelines on AI and LLMs in health research were also identified.^{38,52-56} These include extensions of existing standards (PRISMA-AI,³⁸ TRIPOD+AI,⁵³ and TRIPOD-LLM⁵²), as well as consensus-based checklists focused more broadly on ML (PALISADE⁵⁵ and REFORMS⁵⁴). These guidelines informed the development of ELEVATE-GenAI by highlighting principles such as model transparency, reproducibility, structured human evaluation, and ethical AI practices. In May 2025, the DEAL checklist was published and will be included in future iterations of the ELEVATE-GenAI framework.⁶⁴

Domain Identification for the ELEVATE-GenAI Framework

The ELEVATE-GenAI framework builds on domains by Bedi et al⁴⁰ and the Holistic Evaluation of Language Models benchmark,⁴⁵ which provide strong foundations for evaluating AI performance. The ISPOR Working Group on Generative AI expanded this structure with 3 additional domains—model characteristics, reproducibility and generalizability, and security and privacy—to address HEOR-specific methodological and regulatory needs. These additions were informed by expert input and gaps identified in the literature review. To assess alignment, components

Figure 1. Inclusion of ELEVATE-GenAI domains across 30 studies and report. Each reference was scored across the 10 ELEVATE-GenAI reporting domains based on whether they were clearly included (score = 2), partially included (score = 1), or not reported (score = 0). The stacked bars show the number of references (N=30) receiving each score within each domain, illustrating variation in inclusion of the ELEVATE-GenAI domains across these studies.



from each reviewed study were mapped to the 10 domains. [Figure 1](#) shows their frequency of inclusion across 30 studies, with accuracy, fairness and bias, and reproducibility and generalizability most frequently addressed, and security and privacy least represented.

Reporting Domains: Definitions and Guidance

The ELEVATE-GenAI reporting guidelines are designed for HEOR studies in which generative AI plays a substantive role in evidence generation, synthesis, or analysis. They are not intended for studies using AI only for minor tasks such as text editing. The 10-domain checklist covers foundational model characteristics (eg, architecture, training data, and access) and output quality across key HEOR applications, such as SLRs, HEM, and RWE ([Table 1](#)). Each domain includes targeted reporting items to help authors clearly describe their use of generative AI, supporting transparency and research integrity. Users should apply judgment in selecting relevant domains and briefly justify any exclusions, allowing flexibility for diverse and evolving HEOR use cases.

To support interpretation, each domain is assigned a maturity level reflecting the current availability of established metrics or reporting standards. High-maturity domains have well-defined practices, whereas low-maturity ones indicate evolving methods. These expert-assessed ratings within the ISPOR Working Group on Generative AI are a pilot feature and will be revisited in future validation. [Table 2](#) outlines the 10 reporting domains and their definitions.

Model Characteristics

This domain focuses on documenting the foundational attributes of the LLM used in the study. Key elements include the model's name (eg, LLaMA-3), version, developer or organization, release date, license type (eg, commercial or open source), and access method (eg, application programming interface [API], web

interface, or local deployment). Authors should also report the model's architecture (eg, transformer based) and provide details about training data sources, where applicable. This includes general-purpose pretraining corpora (where identifiable), data sets used for fine-tuning or instruction-tuning, any proprietary data used for custom models, and any sources integrated into retrieval-augmented generation workflows. Where applicable, authors are encouraged to discuss the explainability of the model's outputs, particularly in relation to interpreting findings in HEOR contexts. Although explainability is not designated as a standalone domain in ELEVATE-GenAI, it remains an important consideration for transparency, reproducibility and stakeholder trust.

Level of Maturity: High

Well-established practices exist for describing model provenance, architecture, and access, although transparency about training data remains limited in some proprietary models.

Accuracy Assessment

This domain evaluates how well Gen AI-generated outputs align with correct or expected results. Accuracy can be assessed through comparisons with human benchmarks, gold-standard data sets, or expert review. Metrics may include commonly used measures in AI/ML, such as precision, recall, F1 score, and area under the curve, as well as natural-language-processing-specific (eg, BLEU) or domain-specific metrics (eg, Generative Radiology Report Evaluation and Error Notation for radiology report generation).⁶⁵ In HEOR, appropriate methods include fact checking against source documents, expert review, or benchmarking against known evidence, but the suitability of accuracy metrics depends on the task. Structured tasks, such as data extraction or classification, lend themselves to quantitative metrics, whereas free-text generation, such as drafting an HTA dossier, often requires qualitative assessment. Although interest is growing in

Table 1. ELEVATE-GenAI checklist for evaluating LLM use in HEOR research.

Model characteristics
Is the model's name, version, developer, release date, license (eg, open source or commercial), and architecture described? Are the training data sources (eg, domain-specific data sets such as PubMed) and fine-tuning details provided?
Accuracy assessment
Are task-specific accuracy metrics (eg, Precision, Recall, and F1 Score) reported, where applicable (accounting for the fact that different metrics will be relevant for different tasks)? Are outputs validated against human benchmarks or gold-standard data sets?
Comprehensiveness assessment
Are outputs compared with relevant benchmarks (eg, published reviews, validated models) to ensure completeness? Is there expert evaluation confirming all critical elements of the task are addressed?
Factuality verification
Are methods for verifying the factual accuracy of outputs (eg, cross-referencing with sources, expert review) described? Are discrepancies and corrective actions documented?
Reproducibility protocols and generalizability
Are reproducibility protocols (eg, training code, query phrasing, and hyperparameters) shared? Are workflows provided to support independent verification? Is the generalizability of the approach and methods to similar research questions addressed?
Robustness checks
Are robustness tests (eg, handling typographical errors, ambiguous queries) documented? Are changes in model performance under these conditions reported?
Fairness and bias monitoring
Are outputs evaluated for biases or stereotypes related to gender, age, ethnicity, or other demographics? Are fairness metrics (eg, demographic parity) used (if applicable), and corrective actions for identified biases documented?
Deployment context and efficiency metrics
Are deployment setup details (eg, hardware, software, and runnable deployment code) clearly described? Are efficiency metrics (eg, processing time, scalability, and resource usage) reported?
Calibration and uncertainty
Are calibration methods (eg, expected calibration error) described (if applicable)? Are thresholds for manual review of outputs (eg, ambiguous cases flagged in systematic reviews) specified?
Security and privacy measures
Are security protocols (eg, encryption, anonymization, and access controls) documented? Is compliance with regulations such as GDPR or HIPAA reported, if applicable? Is compliance with intellectual property and copyright law documented?
Overall score: Assign 3 points for each domain rated as Clearly Reported, 2 points for Ambiguous, and 1 point for Not Reported. Sum the points across all domains to calculate the overall score.

adapting AI/ML accuracy measures for HEOR tasks, such as SLRs and HEM, and in developing HEOR-specific benchmarks, further work is needed to define fit-for-purpose evaluation strategies tailored to these specific contexts.

Level of Maturity: Medium

Core accuracy concepts are well developed in the AI/ML field; however, guidance on HEOR-specific implementation, particularly for text generation tasks, remains limited and evolving.

Comprehensiveness Assessment

This domain focuses on evaluating whether GenAI-generated outputs fully and coherently address all required elements of the assigned task. In the context of HEOR, this may include ensuring that all relevant studies are captured in a systematic review, that all model components and assumptions are described in an economic evaluation, or that all relevant outcomes and perspectives are considered in value assessments. Outputs should be compared with authoritative references, such as established guidelines, benchmark publications, or prior high-quality submissions. Expert review can help determine whether critical elements are missing

or inadequately addressed. Comprehensiveness is distinct from accuracy: although accuracy relates to the correctness of specific elements, comprehensiveness assesses whether all relevant content has been fully and coherently addressed. For example, a meta-analysis may accurately describe included studies yet still be incomplete if it omits a pivotal trial. Ensuring completeness is essential to support informed decision making based on the full body of evidence.

Level of Maturity: High

Although typically assessed qualitatively, there are well-established expectations for comprehensiveness across many HEOR tasks, supported by reporting guidelines and expert standards.

Factuality Verification

This domain focuses on verifying that model-generated outputs are factually correct and supported by reliable sources. In HEOR, this includes confirming the accuracy of cited data, study findings, and modeling assumptions through expert review, cross-checking with primary sources, or automated source attribution

Table 2. An evaluation framework for large-language models focused on evidence, transparency, and efficiency (the ELEVATE-GenAI framework) (adapted from Holistic Evaluation of Language Models and Bedi et al⁴⁰).

Domain name	Domain description	Reporting guidelines	Level of maturity of domain measurement
Model characteristics	Describes the model's foundational characteristics, such as name, version, developer, model access, license, release date, architecture, training data, and fine-tuning performed for specific tasks.	<ul style="list-style-type: none"> - Provide details of the model, including name, version, developer(s), release date, license (eg, commercial or open source), access (eg, links to the models), architecture (eg, transformer based). - Describe training data, including domain-specific sources (eg, PubMed) and any fine-tuning performed. 	High
Accuracy assessment	Measures how closely the model's output aligns with the correct or expected answer, evaluating precision, relevance, and correctness.	<ul style="list-style-type: none"> - Compare results with human benchmarks or gold-standard data sets for validation. - If appropriate for the task at hand, report metrics (eg, Precision, Recall, F1 Score, and AUC). These metrics will not be applicable to all tasks. 	Medium: further work required on adapting AI/ML metrics to HEOR studies and identifying appropriate metrics for specific tasks.
Comprehensiveness assessment	Assesses how thoroughly the model's output addresses all aspects of the task, ensuring completeness, coherence, and critical coverage.	<ul style="list-style-type: none"> - Evaluate completeness by comparing outputs with benchmarks, such as published reviews or models. - Use expert evaluations to confirm critical elements are addressed. 	High
Factuality verification	Evaluates whether the model's output is accurate and based on verifiable sources, identifying hallucinated or non-existent citations.	<ul style="list-style-type: none"> - Explain methods to verify factual accuracy (eg, expert review and source validation). - Document discrepancies and corrective actions taken. 	High
Reproducibility protocols and generalizability	Ensures methods and outputs can be independently verified by documenting workflows, sharing code, and specifying hyperparameters. Evaluates generalizability of approach proposed.	<ul style="list-style-type: none"> - List reproducibility protocols, including training code, query phrasing, and hyperparameters. - Share workflows to facilitate independent verification. - Address generalizability of methods to similar research questions. 	High
Robustness checks	Tests the model's resilience to input variations, such as typographical errors or ambiguous queries.	<ul style="list-style-type: none"> - Document robustness tests, including handling of typos, adversarial inputs, or ambiguous phrasing. - Report any changes in performance under these conditions. 	High

continued on next page

Table 2. Continued

Domain name	Domain description	Reporting guidelines	Level of maturity of domain measurement
Fairness and bias monitoring	Evaluates whether the model's output is equitable and free from harmful biases or stereotypes across diverse groups and contexts.	<ul style="list-style-type: none"> - Monitor fairness by checking for bias in outputs related to gender, age, ethnicity, or other demographics. - If appropriate, use fairness metrics, such as demographic parity and document corrective actions if biases are identified. 	Low: the use of metrics to assess fairness and bias is an ongoing area of research
Deployment context and efficiency metrics	Examines the technical setup, resource requirements, and efficiency metrics to evaluate practical feasibility.	<ul style="list-style-type: none"> - Describe deployment setup, including hardware (eg, NVIDIA A100 GPUs) and software (eg, TensorFlow and PyTorch) and runnable deployment code (eg, via Docker) - Report efficiency metrics, such as processing time, scalability, and resource efficiency. 	High
Calibration and uncertainty	Measures how well the model conveys uncertainty in its outputs, including confidence levels and its ability to handle ambiguity appropriately.	<ul style="list-style-type: none"> - If appropriate for the task at hand, describe calibration methods and metrics appropriate for the task (eg, expected calibration error) - Specify thresholds for flagging outputs requiring manual review (eg, percent of abstracts included in screening in SLR). 	Low: the use of metrics to evaluate calibration and uncertainty is an ongoing area of research
Security and privacy measures	Assesses adherence to security, privacy, and data protection standards and regulations, including anonymization, secure handling, and compliance with regulations such as GDPR or HIPAA, if appropriate.	<ul style="list-style-type: none"> - Describe security protocols, such as data encryption, anonymization, and access controls. - Ensure compliance with regulations such as GDPR or HIPAA if appropriate - Document measures to safeguard intellectual property and copyright. 	Low: identifying the appropriate metrics for this domain is an ongoing area of research
Overall score	Calculates an overall score for the evaluation using the checklist.	Assign 3 points for each domain rated as Clearly Reported, 2 points for Ambiguous, and 1 point for Not Reported. Sum the points across all domains to calculate the overall score.	Low: the usefulness of this score will need to be further evaluated through feedback from the HEOR community

AUC indicates Area under the curve; GDPR, General Data Protection Regulation; GPU, Graphics Processing Unit; HIPAA, Health Insurance Portability and Accountability Act; LLM, large language model.

where available. A key concern is the identification and correction of hallucinated or fabricated content, such as false citations, misrepresented results, or unsupported claims.¹⁹ Authors should document any discrepancies found during review and describe the steps taken to address them. Factuality is distinct from accuracy: although accuracy reflects alignment with expected results or benchmarks, factuality concerns the truthfulness and verifiability of the content itself. For instance, a summary may accurately capture a study's structure but misreport specific findings, resulting in factual errors despite an otherwise accurate format.

These distinctions, although nuanced, are important for ensuring trust in LLM-generated outputs and will be further evaluated during the piloting and validation phases described in this manuscript.

Level of Maturity: High

Established practices for fact checking and source validation are already in place in scientific research workflows and can be readily applied to AI-generated outputs.

Reproducibility Protocols and Generalizability

This domain assesses 2 critical aspects of reliability: reproducibility or the ability to replicate results and generalizability, the applicability of methods across different contexts. Reproducibility is essential for scientific credibility and policy relevance; yet, it can be difficult to achieve in generative AI because of proprietary models, frequent updates, and the stochastic nature of outputs. The dynamic nature of some generative AI systems—particularly those that continuously learn or are regularly updated—further complicates reproducibility. To mitigate these challenges, researchers should document key contextual details, including model version, date of access, deployment method (eg, API or local instance), prompt wording, and relevant system settings (eg, temperature, seed).^{54,66} When full transparency is not possible—especially with commercial or black-box models—authors should clearly state these limitations. Retrieval-augmented generation approaches may enhance reproducibility by grounding model outputs in verifiable sources, providing a potential pathway for more consistent and auditable results across studies.^{67,68}

Generalizability involves assessing whether the LLM workflow can be applied to other HEOR questions, populations, or settings. For narrow or single-use applications, authors should indicate that generalizability does not apply and briefly explain why. Both dimensions help ensure responsible, scalable use of LLMs in HEOR.

Level of Maturity: High

Although some implementation challenges persist, particularly for closed-source systems, reproducibility documentation practices are well established, and generalizability is a routine consideration in HEOR research.

Robustness Checks

This domain focuses on evaluating the model's resilience to variations in input, such as typographical errors, ambiguous phrasing, or minor changes in prompt structure. In HEOR applications, this may be particularly important for tasks that rely on consistent and interpretable output (eg, data extraction or structured summarization). Authors should report whether robustness testing was performed and describe any observed variation in output quality or performance under perturbed input conditions. In cases in which inputs and prompts are fully standardized and tightly constrained, such as in highly scripted workflows or API-based automations, robustness checks may be less relevant. Authors should briefly note when robustness testing was not conducted and explain why it was not applicable.

Level of Maturity: High

Robustness testing is widely recognized in AI/ML research and is increasingly incorporated into evaluation practices for LLM applications in health and biomedical research.

Fairness and Bias

This domain focuses on identifying and mitigating potential biases in model-generated outputs to ensure equity across populations and avoid reinforcing harmful stereotypes or exclusions. In the HEOR context, fairness may relate to how outputs differ across sociodemographic groups, such as gender, age, ethnicity, or socioeconomic status. Where applicable, authors are encouraged to assess fairness using established metrics, such as demographic parity or equalized odds, and to evaluate output consistency across relevant subgroups.⁶⁹⁻⁷¹ However, this remains an area of active methodological development, and selecting appropriate fairness metrics and implementing subgroup analyses may require

specialized expertise, particularly in HEOR applications. Authors should indicate whether fairness or bias assessments were conducted and describe any relevant findings. If this domain is not applicable to the study (eg, if the LLM is not generating person-level or subgroup-relevant content), authors should briefly explain why it was excluded.

Level of Maturity: Low

Although fairness is a critical consideration, practical guidance and validated metrics for generative AI in HEOR remain limited and evolving.

Deployment Context and Efficiency Metrics

This domain addresses both the technical configuration of the model deployment and the efficiency of its operation. Authors should describe the deployment setup, including hardware specifications (eg, number and type of graphical processing units [GPUs], such as NVIDIA A100, H100, or tensor processing unit variants), software frameworks (eg, Hugging Face Transformers) and orchestration tools (eg, Docker and Ray). When possible, authors should indicate whether deployment artifacts, such as container images, configuration files, environment specifications, or API wrappers, are publicly available to facilitate reproducibility. Efficiency metrics are also essential for assessing the model's scalability and practical utility in HEOR applications. Relevant metrics may include latency (response time per query), throughput (eg, documents processed per second), compute efficiency (eg, floating-point operations per token), and cost metrics (eg, token cost for commercial APIs). For example, time and cost required to generate outputs for tasks such as SLRs or HEMs may significantly influence feasibility of large-scale deployment. When models are accessed via APIs (eg, commercial models, such as GPT-4o), efficiency considerations should also include token limits, response latency, usage costs, and rate limits, all of which may affect scalability, reproducibility, and real-world applicability.

Level of Maturity: High

Clear practices exist for reporting deployment configurations and performance metrics, especially for reproducible research and cloud-based applications.

Calibration and Uncertainty

This domain evaluates whether the model expresses uncertainty appropriately and whether its confidence aligns with actual performance. Calibration is particularly important in HEOR, in which overconfidence or underconfidence in outputs can lead to misinformed decisions. Metrics such as expected calibration error (ECE)⁷² are being explored for HEOR use but remain underdeveloped. In SLRs, for instance, uncertainty thresholds can help flag abstracts for manual review as part of hybrid AI-human workflows.⁴⁵ However, such metrics are not yet widely adopted in HEOR and require further validation. Authors should report whether uncertainty was assessed, how it was quantified, and whether the model's confidence appeared well calibrated for the task. If this domain is not applicable—eg, for tasks where confidence estimation is not used—authors should state this and provide a brief justification.

Level of Maturity: Low

Although the concept of calibration is well defined in AI/ML, practical tools and norms for uncertainty quantification in HEOR applications remain limited and evolving.

Table 3. Application of the ELEVATE-GenAI checklist to a systematic literature review study (Robinson et al).¹⁶

Checklist questions	Domain evaluation	Assessment
<p>1. Model characteristics</p> <p>Is the model's name, version, developer, release date, license (eg, open source or commercial), and architecture described? Are the training data sources (eg, domain-specific data sets such as PubMed) and fine-tuning details provided?</p>	<p>The Bio-SIEVE model is based on instruction-tuned versions of LLaMA7B and Guanaco7B, using a 7B parameter architecture with quantization (4-bit). BIO-SIEVE is not open source, although several elements (eg, code and parameters) are provided. The publication date is 2023. Training involved 7330 systematic reviews from Cochrane, focusing on inclusion/exclusion criteria and reasoning for abstract exclusion. Instruction fine-tuning was conducted to improve performance on systematic review tasks.</p>	<p>Clearly Reported</p> <p>This item was rated as Clearly Reported because the model name, architecture, developer, license status, training sources, and fine-tuning procedures were all described in detail, including the use of Cochrane data sets and task-specific tuning.</p>
<p>2. Accuracy assessment</p> <p>Are task-specific accuracy metrics (eg, Precision, Recall, and F1 Score) reported where applicable (accounting for the fact that different metrics will be relevant for different tasks)? Are outputs validated against human benchmarks or gold-standard data sets?</p>	<p>The article reports precision, recall, and accuracy metrics for inclusion/exclusion tasks, comparing Bio-SIEVE's performance with baseline models (eg, logistic regression) and other LLMs such as ChatGPT. Bio-SIEVE achieved higher recall and accuracy for inclusion/exclusion but underperformed in exclusion reasoning, where ChatGPT demonstrated better results.</p>	<p>Clearly Reported</p> <p>This item was rated as Clearly Reported because precision, recall, and accuracy metrics were reported and benchmarked against human labels and multiple baselines, including other LLMs.</p>
<p>3. Comprehensiveness assessment</p> <p>Are outputs compared with relevant benchmarks (eg, published reviews and validated models) to ensure completeness? Is there expert evaluation confirming all critical elements of the task are addressed?</p>	<p>Bio-SIEVE's outputs were validated against gold-standard data sets (eg, Cochrane) and expert-annotated safety-first subsets. The Bio-SIEVE Guanaco7B (Single) achieved a precision of 0.85 and a recall of 0.82 on the test set, demonstrating a strong balance between minimizing false positives and capturing relevant abstracts (but performed less well on the safety-first subset). Expert validation confirmed no critical gaps in inclusion, aligning with the goal of capturing all potentially relevant abstracts during screening.</p>	<p>Clearly Reported</p> <p>This item was rated as Clearly Reported because outputs were benchmarked against gold-standard data sets, and expert validation confirmed no critical gaps in inclusion coverage</p>
<p>4. Factuality verification</p> <p>Are methods for verifying the factual accuracy of outputs (eg, cross-referencing with sources and expert review) described? Are discrepancies and corrective actions documented?</p>	<p>Exclusion reasoning and inclusion/exclusion decisions were cross-referenced with expert-annotated data sets. Discrepancies (eg, missed inclusions) were documented and analyzed, with manual reviews of ambiguous cases ensuring factual reliability.</p>	<p>Clearly Reported</p> <p>This item was rated as Clearly Reported because inclusion/exclusion outputs were compared with expert-annotated references, and discrepancies were documented and manually reviewed.</p>
<p>5. Reproducibility protocols and generalizability</p> <p>Are reproducibility protocols (eg, training code, query phrasing, and hyperparameters) shared? Are workflows provided to support independent verification? Is the generalizability of the approach and methods to similar research questions addressed?</p>	<p>Detailed reproducibility information includes fine-tuning parameters (eg, batch size and learning rate), preprocessing workflows, and access to training data sets. Access to code and adapter weights is provided on HuggingFace. The approach is generalizable to abstract screening tasks in other medical domains.</p>	<p>Clearly Reported</p> <p>This item was rated as Clearly Reported because training parameters, code, data sets, and model adapters were shared, and generalizability to other medical domains was addressed.</p>

continued on next page

Table 3. Continued

Checklist questions	Domain evaluation	Assessment
6. Robustness checks		Clearly Reported
Are robustness tests (eg, handling typographical errors and ambiguous queries) documented? Are changes in model performance under these conditions reported?	Robustness was tested by varying input prompts and testing irrelevancy exclusions (eg, pairing abstracts with unrelated topics). Bio-SIEVE consistently excluded irrelevant abstracts, demonstrating robustness to input variations.	This item was rated as Clearly Reported because robustness was tested by varying inputs and pairing abstracts with unrelated content, demonstrating consistent model behavior.
7. Fairness and bias monitoring		Not Reported
Are outputs evaluated for biases or stereotypes related to gender, age, ethnicity, or other demographics? Are fairness metrics (eg, demographic parity) used (if applicable), and corrective actions for identified biases documented?	Fairness metrics, such as demographic parity, or bias in inclusion/exclusion decisions, were not explicitly assessed. Population biases were not evaluated.	This item was rated as Not Reported because no analysis of demographic or representational bias was conducted, and fairness metrics were not applied.
8. Deployment context and metrics		Ambiguous
Are deployment setup details (eg, hardware, software, and runnable deployment code) clearly described? Are efficiency metrics (eg, processing time, scalability, and resource usage) reported?	The Bio-SIEVE Guanaco7B models were trained on 4 NVIDIA A100 80GB GPUs for 24-40 hours, depending on the model. Inference time was reported as 1.39 seconds per sample on an RTX 3090 GPU, but context (eg, batch size) and memory usage metrics were not provided.	This item was rated as Ambiguous because GPU usage and inference time were reported, but key efficiency metrics such as batch size, memory consumption, and scalability were not provided.
9. Calibration and uncertainty		Ambiguous
Is the model's uncertainty quantified and explicitly reported (if applicable)? Are thresholds for manual review of outputs (eg, ambiguous cases flagged in systematic reviews) specified?	Confidence in inclusion/exclusion decisions was not explicitly quantified. Manual validation of safety-first decisions suggests effective uncertainty management, but explicit thresholds were not defined.	This item was rated as Ambiguous because confidence levels and thresholds for ambiguity were not quantified, although manual validation suggests some awareness of uncertainty.
10. Security and privacy measures		Not Reported
Are security protocols (eg, encryption, anonymization, and access controls) documented? Is compliance with regulations such as GDPR or HIPAA reported, if applicable? Is compliance with intellectual property and copyright law documented?	Compliance with AI regulations, copyright protection, and data security were not discussed. Patient-level data were not used, minimizing direct privacy risks.	This item was rated as Not Reported because security, privacy, and regulatory compliance were not discussed, although the study avoided using identifiable patient data.
Overall score: Assign 3 points for each domain rated as Clearly Reported, 2 points for Ambiguous, and 1 point for Not Reported. Sum the points across all domains to calculate the overall score.		Clearly Reported: 6, Ambiguous: 2, Not Reported: 2 Total score = 24/30

Note. The scoring system (3 = Clearly Reported, 3 = Not Applicable, 2 = Ambiguous, and 1 = Not Reported) is optional and intended only for the self-assessment of reporting completeness. It does not reflect methodological rigor or study quality. The scoring system will be piloted and reassessed in future validation rounds. GPU indicates graphics processing unit; LLM, large language model.

Security and Privacy

This domain evaluates whether appropriate safeguards are in place to protect sensitive, personal, or proprietary data used during model development or output generation. In HEOR studies that involve personal health information, clinical records, or licensed content, authors should describe relevant security protocols, including encryption methods, anonymization techniques, and access controls. Where applicable, authors should also indicate whether their work complies with data protection regulations, such as General Data Protection Regulation (GDPR) or Health Insurance Portability and Accountability Act (HIPAA), and describe any measures taken to protect intellectual property or copyrighted material.³ Security and privacy protections are essential to maintaining stakeholder trust,

regulatory compliance, and research integrity. If the study does not involve sensitive or proprietary data, authors may state that this domain is not applicable and provide a brief explanation.

Level of Maturity: Low

Although security and privacy principles are well established in healthcare and technology, specific implementation guidance for generative AI use in HEOR is still emerging.

Overall Score (Optional)

The scoring system is an optional tool to help users and reviewers assess the completeness of reporting. It is not a required domain and is not needed for framework adherence. Each domain

Table 4. Application of the ELEVATE-GenAI checklist to a health economic modeling study (Reason et al²³).

Checklist questions	Evaluation	Assessment
1. Model characteristics Is the model's name, version, developer, release date, license (eg, open source or commercial), and architecture described? Are the training data sources (eg, domain-specific data sets such as PubMed) and fine-tuning details provided?	The study utilized GPT-4, a transformer-based large language model developed by OpenAI, a commercial model. Specific GPT-4 model release date was not specified. The model was accessed via API, and no specific fine-tuning for health economic modeling was reported. GPT-4 training data includes general-purpose data sets. Explicit adaptation for health economic modeling tasks was absent. In this study, domain-specific functionality was achieved through iterative development of contextual prompts.	Ambiguous This item was rated as Ambiguous because some key elements—such as the model release date, fine-tuning details, and use of domain-specific training data—were not reported, although general model characteristics and access method were described.
2. Accuracy assessment Are task-specific accuracy metrics (eg, Precision, Recall, and F1 Score) reported where applicable (accounting for the fact that different metrics will be relevant for different tasks)? Are outputs validated against human benchmarks or gold-standard data sets?	Accuracy was assessed by comparing model outputs with the published model results. For NSCLC, 93% of runs were completely error-free; for RCC, 60% of runs required simplification but were error-free. ICERs were within 1% of published values.	Clearly Reported This item was rated as Clearly Reported because model outputs were quantitatively compared with published benchmarks, and error rates and ICER deviations were clearly documented.
3. Comprehensiveness assessment Are outputs compared with relevant benchmarks (eg, published reviews and validated models) to ensure completeness? Is there expert evaluation confirming all critical elements of the task are addressed?	Outputs replicated complete three-state models, including progression-free, progressed disease, and death states. Simplification of complex RCC model steps was noted. Benchmarking against published results ensured alignment.	Clearly Reported This item was rated as Clearly Reported because the outputs included all key model components and were benchmarked against complete published models, with expert interpretation noted.
4. Factuality verification Are methods for verifying the factual accuracy of outputs (eg, cross-referencing with sources and expert review) described? Are discrepancies and corrective actions documented?	ICERs and transition values were cross-referenced with published models. Minor discrepancies (eg, discounting assumptions) were documented and attributed to differences in software calculation methods.	Clearly Reported This item was rated as Clearly Reported because the model outputs were cross-checked against source materials, discrepancies were noted, and explanations were provided.
5. Reproducibility protocols and generalizability Are reproducibility protocols (eg, training code, query phrasing, and hyperparameters) shared? Are workflows provided to support independent verification? Is the generalizability of the approach methods to similar research questions addressed?	Prompts, API parameters, and Python-based automation workflows were described, enabling reproducibility. Generated R scripts are publicly available for independent validation. Prompting strategies for the NSCLC model were re-used for the RCC model without modification suggesting their potential applicability across different health economic decision problems.	Clearly Reported This item was rated as Clearly Reported because detailed prompts, parameters, and automation scripts were shared, and the reuse of prompt strategies across models supported generalizability.
6. Robustness checks Are robustness tests (eg, handling typographical errors and ambiguous queries) documented? Are changes in model performance under these conditions reported?	Robustness was tested through prompt variation, such as breaking scripts into multiple prompts. Simplifications were required for overly complex RCC calculations, demonstrating some limitations in handling atypical scenarios.	Clearly Reported This item was rated as Clearly Reported because prompt variations were tested, and limitations in handling complex inputs were described and interpreted in context.

continued on next page

Table 4. Continued

Checklist questions	Evaluation	Assessment
7. Fairness and bias monitoring		Not Reported
Are outputs evaluated for biases or stereotypes related to gender, age, ethnicity, or other demographics? Are fairness metrics (eg, demographic parity) used (if applicable), and corrective actions for identified biases documented?	The study did not explicitly address fairness or demographic bias. Outputs were focused on technical replication of published models without discussion of bias or fairness in population representation.	This item was rated as Not Reported because there was no assessment of fairness or bias related to demographic factors, nor any mention of mitigation strategies.
8. Deployment context and metrics		Clearly Reported
Are deployment setup details (eg, hardware, software, and runnable deployment code) clearly described? Are efficiency metrics (eg, processing time, scalability, and resource usage) reported?	Deployment used Python and R, with scripts processed on midrange GPUs. Average generation times were 715 seconds for the NSCLC model and 956 seconds for the RCC model. Automation using Python streamlined interactions with GPT-4, improving scalability for larger data sets by reducing manual intervention. Time to create context-specific prompts was not reported.	This item was rated as Clearly Reported because the computational setup and processing time were described, along with the use of automation to improve scalability.
9. Calibration and uncertainty		Not Reported
Is the model's uncertainty quantified and explicitly reported (if applicable)? Are thresholds for manual review of outputs (eg, ambiguous cases flagged in systematic reviews) specified?	Model outputs varied slightly across 15 runs, despite low-temperature settings. Manual quality assurance flagged errors and confirmed minor variability in ICERs. Explicit uncertainty quantification was not performed.	This item was rated as Not Reported because uncertainty quantification was not performed, and there were no defined thresholds or formal handling of ambiguous outputs.
10. Security and privacy measures		Clearly Reported
Are security protocols (eg, encryption, anonymization, and access controls) documented? Is compliance with regulations such as GDPR or HIPAA reported, if applicable? Is compliance with intellectual property and copyright law documented?	Dummy data replaced sensitive inputs in prompts because of concerns about LLM data retention. The paper suggests private LLM instances as a future solution to address security and intellectual property concerns.	This item was rated as Clearly Reported because data protection strategies were described, including the use of dummy inputs and future recommendations for secure deployment.
Overall score: Assign 3 points for each domain rated as Clearly Reported, 2 points for Ambiguous, and 1 point for Not Reported. Sum the points across all domains to calculate the overall score.		Clearly Reported: 7, Ambiguous: 1, Not Reported: 2 Total score: 25/30
<p><i>Note.</i> The scoring system (3 = Clearly Reported, 3 = Not Applicable, 2 = Ambiguous, and 1 = Not Reported) is optional and intended only for the self-assessment of reporting completeness. It does not reflect methodological rigor or study quality. The scoring system will be piloted and reassessed in future validation rounds. API indicates application programming interface; ECE, expected calibration error; ICER, incremental cost-effectiveness ratio; LLM, large language model; NSCLC, non-small cell lung cancer; RCC, renal cell carcinoma.</p>		

can be rated on a 3-point scale: Clearly Reported (3 points), Not Applicable (3 points), Ambiguous (2 points), or Not Reported (1 point). “Clearly Reported” indicates full adherence to domain criteria; “Not Applicable” reflects domains irrelevant to the study; “Ambiguous” refers to incomplete or unclear reporting; and “Not Reported” means relevant information is missing. The total score, calculated by summing across domains, offers a summary of reporting completeness and may support self-assessment or peer review. However, it should not be interpreted as a measure of methodological rigor. The scoring feature is optional and designed to support consistent reporting—not to grade or rank studies. Alternative approaches, such as flagging missing critical domains, will be explored in future iterations of the framework.

Level of Maturity: Low

Although scoring systems are common in reporting guidelines, their application to LLM use in HEOR is still under development and requires further testing.

Applications of the ELEVATE-GenAI Framework to HEOR Activities

The ELEVATE-GenAI reporting framework was applied to 2 published HEOR use cases to illustrate its applicability: one focused on abstract screening for a systematic literature review (SLR)¹⁶ and the other on developing a cost-effectiveness model for health economic analysis.²³ These examples, detailed in Tables 3¹⁶ and 4,²³ illustrate how the framework can be used to systematically assess the reporting of outputs augmented with LLMs across distinct HEOR tasks.

ELEVATE-GenAI Application to an SLR Publication

Table 3¹⁶ shows the application of the ELEVATE-GenAI framework to evaluate the Bio-SIEVE model in the SLR study by Robinson et al.¹⁶ This study investigates the use of LLMs to automate title and abstract screening for SLR in the biomedical field and assesses the performance of LLMs in exclusion reasoning, (ie,

providing the rationale for excluding an abstract). The model, instruction tuned on LLaMA and Guanaco, uses a 7B parameter architecture with quantization (4-bit LoRA) and was trained on 7330 Cochrane systematic reviews, focusing on inclusion/exclusion criteria. Fine-tuning was validated with a curated safety-first test set to ensure task-specific performance. Accuracy metrics, such as precision, recall, and overall accuracy, demonstrated superior performance compared with logistic regression and other LLMs (eg, ChatGPT). Comprehensiveness was validated against gold-standard data sets and expert reviews to ensure no relevant abstracts were missed. Factuality verification involved cross-checking inclusion/exclusion decisions with expert data sets, with discrepancies documented and addressed. Reproducibility protocols included detailed documentation of fine-tuning parameters and workflows, with publicly available code and weights for independent validation. The methods are likely generalizable to other medical domains. Robustness was assessed by varying input prompts, with Bio-SIEVE consistently excluding irrelevant abstracts. Fairness and bias monitoring were not explicitly measured. Deployment metrics, including hardware specifications (eg, 4 A100 GPUs) and processing time (eg, 1.39 seconds per sample), highlighted scalability. Calibration and uncertainty measures were limited, relying on manual validation without explicit thresholds for ambiguous cases. Security and privacy were addressed through anonymization and secure handling of Cochrane data, but copyright protection was not discussed. Compliance with HIPAA or GDPR would not be relevant to this type of study.

In summary, the application of Bio-SIEVE study by Robinson et al¹⁶ found that 6 domains were “Clearly Reported,” 2 were “Ambiguous,” and 2 were “Not Reported.” As expected, 3 out of the 4 domains that were evaluated as ambiguous or not reported (fairness and bias monitoring, calibration and uncertainty, and security and privacy measures) correspond to domains with a low level of maturity for metrics, further highlighting the need for future work to identify the useful metrics for these domains.

Application to an HEM Publication

Table 4²³ demonstrates the application of the ELEVATE-GenAI framework to an HEM study by Reason et al.²³ The study explores the feasibility of using GPT-4 to automatically program health economic models. Specifically, the study aims to replicate 2 existing health economic analyses: the cost-effectiveness of nivolumab versus docetaxel for non-small cell lung cancer (NSCLC) and nivolumab plus ipilimumab versus sunitinib and pazopanib for renal cell carcinoma (RCC). The authors provided a detailed description of GPT-4, the LLM used in their study. Accuracy was demonstrated by replicating published 3-state models (progression-free, progressed disease, and death states) with outputs aligning closely to benchmark results, as assessed by comparing incremental cost-effectiveness ratios (ICERs) with published values. For NSCLC models, 93% of runs were error-free, whereas RCC models required simplification but still achieved accuracy within 1% of published ICERs. Precision and recall metrics are not applicable to this use case. Comprehensiveness was validated through benchmarking and replication of complete models, although the need to simplify complex RCC calculations highlighted some limitations. Factuality verification cross-referenced ICERs and transition values with published sources, with minor discrepancies attributed to differences in discounting methods. Reproducibility was supported by detailed prompts, API parameters, and automation workflows, with generated R scripts made publicly available. Generalizability was demonstrated by the successful reuse of prompting strategies from the NSCLC model in the

RCC model without modification, suggesting their potential applicability across different health economic decision problems. Robustness was tested by varying prompts, revealing limitations in handling atypical scenarios, such as overly complex calculations for RCC. Fairness was not explicitly addressed because the study focused on technical replication rather than equity considerations. Deployment relied on Python and R scripts processed on midrange GPUs, with generation times averaging 715 seconds for NSCLC and 956 seconds for RCC. Scalability was improved through automation workflows. Calibration and uncertainty were evaluated qualitatively, with minor ICER variability noted across runs. Security and privacy were addressed by using dummy data to replace sensitive inputs, and the authors suggested private LLM instances as a future solution to enhance security and intellectual property protections.

The HEM study by Reason et al²³ effectively demonstrated the use of LLMs in cost-effectiveness modeling but omitted information required for several domains in the ELEVATE-GenAI framework. The evaluation found that 7 domains were “Clearly Reported,” 1 was “Ambiguous,” and 2 were “Not Reported.” One of the domains, Model Characteristics, was evaluated as Ambiguous, but it would not be difficult for the authors in further iterations to report the appropriate information for this domain, indicating why the ELEVATE-GenAI framework has an important role to play in standardizing what authors might report.

Discussion

Limitations of the ELEVATE-GenAI Reporting Guidelines

The ELEVATE-GenAI guidelines provide a foundational framework for reporting LLM use in HEOR, but several limitations should be acknowledged. First, the targeted literature review informing the framework was not systematic and may have omitted relevant sources. The 10 domains were derived through expert consensus and literature synthesis, but further validation is needed to ensure all relevant aspects of LLM use in HEOR are captured without introducing unnecessary complexity and reporting burden. Maturity levels for each domain reflect expert judgment and are inherently subjective; their value will need to be tested through stakeholder feedback. Similarly, although a scoring system was piloted to support self-assessment, its future utility will depend on broader user input.

Second, certain domain definitions may be challenging to apply consistently because they are conceptually similar. For example, distinguishing between accuracy and comprehensiveness is not always straightforward—an LLM may correctly report included studies (accuracy) but fail to capture all relevant ones (comprehensiveness). Reproducibility is also difficult to achieve, given variability in data access, prompt design, and computational environments. Even with open-source models, exact replication may not be possible, and closed-source models, such as GPT-4, introduce further uncertainty due to ongoing updates.

Third, the framework’s generalizability across HEOR tasks requires further empirical testing (Appendix 3 in Supplemental Materials found at <https://doi.org/10.1016/j.jval.2025.06.018>). Although designed to be broadly applicable, it has only been applied to 2 use cases. Because it is tested across a wider range of activities—such as SLRs, HEM, and RWE generation—its strengths and limitations will become clearer.

Fourth, many evaluation metrics commonly used in AI/ML—such as expected calibration error (ECE), robustness and accuracy metrics—have not been validated for HEOR-specific tasks, such as parameter estimation or health state identification. Fairness and bias assessment remain particularly challenging, especially in the

context of HEOR studies. Of note, benchmarks specific to HEOR field are needed. One example might be a benchmark to evaluate the accuracy of a LLM to screen titles and abstracts in an SLR. To signal the variability in metric maturity, the guidelines assign a “level of maturity” to each domain. Future work should prioritize adapting these metrics to HEOR, refining reporting guidance.

Finally, as agentic approaches become more prevalent—in which LLMs perform iterative or semiautonomous tasks—future versions of ELEVATE-GenAI may require additional guidance in this area.

Next Steps

This version of the ELEVATE-GenAI reporting guidelines was developed through expert input and a targeted literature review. Revisions to date have clarified that scoring is optional, acknowledged the absence of a standalone explainability domain, and recognized that not all domains will apply to every use case. As a living guideline, future versions will be publicly released with opportunities for community input. Next steps could include structured stakeholder consultation, piloting across a range of HEOR applications, and a formal Delphi process to assess the relevance, clarity, and utility of each domain. These activities—modeled after best practices from guideline initiatives such as PRISMA-AI³⁸—will ensure the framework remains practical, flexible, and responsive to the evolving landscape of generative AI in HEOR.

Conclusions

As the use of generative AI accelerates within HEOR, there is an urgent need for rigorous, consistent, and transparent reporting practices. LLMs offer promising capabilities to support evidence generation across tasks such as SLRs, economic modeling, and real-world data analysis. The ELEVATE-GenAI reporting guidelines provide a structured approach for documenting both model characteristics and output quality, helping to ensure scientific integrity in AI-augmented research. Initial applications of the guidelines have identified important areas for refinement, particularly around reproducibility, robustness, fairness, and uncertainty. As generative AI continues to evolve, so too must the tools used to guide its responsible integration into HEOR workflows. By adopting and iteratively improving structured reporting practices, the HEOR community can advance innovation while upholding standards of transparency and trustworthiness.

Glossary

- **Artificial intelligence (AI):** A broad field of computer science that aims to create intelligent machines capable of performing tasks typically requiring human intelligence.
- **Area under the curve (AUC):** A performance metric for classification models that measures the ability to distinguish between classes. It represents the area under the Receiver Operating Characteristic curve, summarizing the trade-off between sensitivity (recall) and specificity. A higher AUC indicates better model performance.
- **Deep learning:** A subset of machine learning algorithms that uses multilayered neural networks, called deep neural networks. These algorithms are the core behind the majority of advanced AI models.
- **Expected calibration error (ECE):** A metric that evaluates how well a model's predicted probabilities align with the actual likelihood of an event occurring. Low ECE indicates better-calibrated predictions, which is critical for applications requiring reliable confidence scores.

- **F1 score:** A metric that balances precision and recall, calculated as the harmonic mean of these 2 measures. It is particularly useful for evaluating models in scenarios where false positives and false negatives have unequal consequences.
- **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 pre-trained transformer (GPT):** A type of large language model (LLM) based on the Transformer architecture, pre-trained on large text data sets to generate human-like language. Although GPT commonly refers to OpenAI's model series (eg, GPT-4), the term also describes a broader class of transformer-based models developed by other organizations, such as Anthropic's Claude.
- **Large language model (LLM):** A specific type of foundation model 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.
- **Precision:** A metric that evaluates the proportion of true positive predictions among all positive predictions made by a model. High precision indicates fewer false positives, which is essential in tasks where accuracy of positive classifications is critical.
- **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 outputs.
- **Recall:** A metric that evaluates the proportion of true positive predictions among all actual positive cases. High recall indicates fewer false negatives, which is crucial for tasks where capturing all relevant instances is a priority.
- **Token:** A token refers to a unit of input data used by a model, which may be a word fragment, symbol, or, in the case of multimodal models, a non-text element such as an image embedding. The context window defines the maximum number of tokens a model can process at once and determines the length and complexity of input it can handle efficiently.

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. Drs Florence, 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.2025.06.018>.

Article and Author Information

Accepted for Publication: June 27, 2025

Published Online: August 20, 2025

doi: <https://doi.org/10.1016/j.jval.2025.06.018>

Author Affiliations: Value Analytics Labs, Cambridge, MA, USA (Fleurence, Ayer); Office of the Director, National Institutes of Health, National Institute of Biomedical Imaging and Bioengineering, Bethesda, MD, USA (Fleurence); National Institute for Health and Care Excellence, London, UK (Dawoud); Cairo University, Faculty of Pharmacy, Cairo, Egypt (Dawoud); 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); Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, USA (Wang); Intelligent Medical Objects, Rosemont, IL, USA (Wang); Institute Department of Biomedical Informatics and Data Science, School of Medicine, Yale University, New Haven, CT, USA (Xu); 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); Center for Health & Humanitarian Systems, Georgia Institute of Technology, Atlanta, GA, USA (Ayer).

Correspondence: Rachael L. Fleurence, PhD, Value Analytics Labs, 100 Cambridge St Suite 1400, Boston, MA 02114, USA. Email: rfleurence@valueanalyticslabs.com

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 program 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 wish to thank the ISPOR Science office for their support and Sahar Alam for her excellent program management throughout the project. Dr Fleurence is a former employee of the National Institutes of Health.

REFERENCES

- Howell MD, Corrado GS, DeSalvo KB. Three epochs of artificial intelligence in health care. *JAMA*. 2024;331(3):242–244.
- Jurafsky D, Martin JH. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd edition. Online manuscript released January 12, 2025. <https://web.stanford.edu/~jurafsky/slp3/>; 2025. Accessed August 13, 2025.
- 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*. 2025;28(2):175–183.
- Telenti A, Auli M, Hie BL, Maher C, Saria S, Ioannidis JPA. Large language models for science and medicine. *Eur J Clin Invest*. 2024:e14183.
- Zhao WX, Zhou K, Li J, et al. A survey of large language models. *arXiv*. published online March 31, 2023. <https://doi.org/10.48550/arXiv.2303.18223>.
- Fleurence R, Wang X, Bian J, et al. Generative AI in health economics and outcomes research: a taxonomy of key definitions and emerging applications, an ISPOR working group report. *Value in Health*. published online February 22, 2025. <https://doi.org/10.48550/arXiv.2410.20204>.
- Open AI, Achiam J, Adler S, et al. GPT-4 technical report. *arXiv*. published online March 4, 2024. <https://doi.org/10.48550/arXiv.2303.08774>.
- Reason T, Klijn S, Rawlinson W, et al. Using generative artificial intelligence in health economics and outcomes research: a primer on techniques and breakthroughs. *Pharmacoecomes*. 2025;9(4):501–517.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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. *Pharmacoecomes*. 2024;8(2):205–220.
- Robinson A, Thorne W, Wu BP, et al. Bio-sieve: Exploring Instruction Tuning Large Language Models for Systematic Review Automation. *arXiv*. published August 12, 2023. <https://doi.org/10.48550/arXiv.2308.06610>.
- 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.
- 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.
- 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.
- Chhatwal J, Samur S, Yildirim IF, Bayraktar E, Ermis T, Ayer T. Fully replicating published health economic models using generative AI. Presented at: ISPOR Europe 2024 Meeting. Barcelona, Spain [https://www.valueinhealthjournal.com/article/S1098-3015\(24\)03392-8/abstract](https://www.valueinhealthjournal.com/article/S1098-3015(24)03392-8/abstract); 2024.
- Chhatwal J, Yildirim IF, Samur S, Bayraktar E, Ermis T, Ayer T. Development of de novo health economic models using generative AI. Presented at: ISPOR Europe 2024 Meeting. Barcelona, Spain [https://www.valueinhealthjournal.com/article/S1098-3015\(24\)02899-7/abstract](https://www.valueinhealthjournal.com/article/S1098-3015(24)02899-7/abstract); 2024.
- Chhatwal J, Yildirim IF, Balta D, et al. Can large language models generate conceptual health economic models? Presented at: ISPOR 2024; 2024; Atlanta, Georgia. <https://www.ispor.org/heor-resources/presentations-database/presentation/intl2024-3898/139128>.
- 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. *Pharmacoecomes*. 2024;8(2):191–203.
- Cohen AB, Waskom M, Adamson B, Kelly J, Amster G. Using large language models to extract PD-L1 testing details from electronic health records. Presented at: ISPOR 2024; 2024; Atlanta, GA <https://www.ispor.org/heor-resources/presentations-database/presentation/intl2024-3898/136019>.
- 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.
- Jiang LY, Liu XC, Nejatian NP, et al. Health system-scale language models are all-purpose prediction engines. *Nature*. 2023;619(7969):357–362.
- 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.
- 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.
- 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.
- Xie Q, Chen Q, Chen A, et al. Me-LLaMa: foundation large language models for medical applications. *Res Sq*. 2024. rs.3.rs-4240043.
- Yang X, Chen A, PourNejatian N, et al. A large language model for electronic health records. *NPJ Digit Med*. 2022;5(1):194.
- US Food and Drug Administration. Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products-guidance for industry and other interested parties. <https://www.fda.gov/media/184830/download>; Published 2025. Accessed August 13, 2025.
- Warraich HJ, Tazbaz T, Califf RM. FDA perspective on the regulation of artificial intelligence in health care and biomedicine. *JAMA*. 2025;333(3):241–247.
- National Institute for Health and Care Excellence. Use of AI in evidence generation: NICE position statement. <https://www.nice.org.uk/about/what-we-do/our-research-work/use-of-ai-in-evidence-generation-nice-position-statement>; Published 2024. Accessed September 20, 2024.
- National Institute for Health and Care Excellence. NICE statement of intent for artificial intelligence (AI). <https://www.nice.org.uk/corporate/ecd12/resources/nice-statement-of-intent-for-artificial-intelligence-ai-pdf-40464270623941>; Published 2024. Accessed December 16, 2024.

36. Canada's Drug Agency. Canada's drug agency position statement on the use of AI in the generation and reporting of evidence. https://www.cda-amc.ca/sites/default/files/MG%20Methods/Position_Statement_AI_Renumbered.pdf; Published 2025. Accessed May 27, 2025.
37. Equator Network. Enhancing the QUALity and transparency of health research. <https://www.equator-network.org>. Accessed May 14, 2025.
38. Cacciamani GE, Chu TN, Sanford DI, et al. Prisma AI reporting guidelines for systematic reviews and meta-analyses on AI in healthcare. *Nat Med*. 2023;29(1):14–15.
39. ALSaad R, Abd-Alrazaq A, Boughorbel S, et al. Multimodal large language models in health care: applications, challenges, and future outlook. *J Med Internet Res*. 2024;26:e59505.
40. 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.
41. Chia YK, Hong P, Bing L, Poria S. Instructeval: towards holistic evaluation of instruction-tuned large language models. *arXiv*. published online June 7, 2023. <https://doi.org/10.48550/arXiv.2306.04757>.
42. de Hond A, Leeuwenberg T, Bartels R, et al. From text to treatment: the crucial role of validation for generative large language models in health care. *Lancet Digit Health*. 2024;6(7):e441–e443.
43. Ko JS, Heo H, Suh CH, Yi J, Shim WH. Adherence of studies on large language models for medical applications published in leading medical journals according to the MI-CLEAR-LLM checklist. *Korean J Radiol*. 2025;26(4):304–312.
44. Lee J, Park S, Shin J, Cho B. Analyzing evaluation methods for large language models in the medical field: a scoping review. *BMC Med Inform Decis Mak*. 2024;24(1):366.
45. Liang P, Bommasani R, Lee T, et al. Holistic evaluation of language models. *arXiv*. published online November 16, 2022. <https://doi.org/10.48550/arXiv.2211.09110>.
46. Moreno AC, Bitterman DS. Toward clinical-grade evaluation of large language models. *Int J Radiat Oncol Biol Phys*. 2024;118(4):916–920.
47. Park SH, Suh CH, Lee JH, Kahn CE, Moy L. Minimum reporting items for clear evaluation of accuracy reports of large language models in healthcare (MI-CLEAR-LLM). *Korean J Radiol*. 2024;25(10):865–868.
48. Park SH, Suh CH. Reporting Guidelines for Artificial Intelligence Studies in Healthcare (for Both Conventional and Large Language Models): what's New in 2024. *Korean J Radiol*. 2024;25(8):687–690.
49. Shi D, Shen T, Huang Y, et al. Large language model safety: a holistic survey. *arXiv*. published online December 23, 2024. <https://doi.org/10.48550/arXiv.2412.17686>.
50. Sun C, Lin K, Wang S, Wu H, Fu C, Wang Z. LalaEval: a holistic human evaluation framework for domain-specific large language models. *arXiv*. published online August 23, 2024. <https://doi.org/10.48550/arXiv.2408.13338>.
51. Wysocka M, Wysocki O, Delmas M, Mutel V, Freitas A. Large Language Models, scientific knowledge and factuality: a framework to streamline human expert evaluation. *J Biomed Inform*. 2024;158:104724.
52. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nat Med*. 2025;31(1):60–69.
53. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378.
54. Kapoor S, Cantrell EM, Peng K, et al. REFORMS: consensus-based recommendations for machine-learning-based science. *Sci Adv*. 2024;10(18):eade3452.
55. 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.
56. Thomas J, Flemmyng E, Noel-Storr A, et al. Responsible AI Evidence Synthesis (RAISE): guidance and recommendations. OSFHOME. <https://osf.io/cn7x4>. Accessed November 26, 2024.
57. Adams L, Fontaine E, Lin S, Crowell T, Chung VCH, Gonzalez A. Artificial intelligence in health, health care, and biomedical science: an AI code of conduct principles and commitments discussion draft. *NAM Perspect*. 2024. <https://doi.org/10.31478/202403a>.
58. Coalition for Health AI. Blueprint for trustworthy AI implementation guidance and assurance for healthcare. https://www.coalitionforhealthai.org/papers/blueprint-for-trustworthy-ai_V1.0.pdf. Accessed April 25, 2023.
59. European Medicines Agency. Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle. <https://www.ema.europa.eu/en/use-artificial-intelligence-ai-medicinal-product-lifecycle>. Accessed May 22, 2024.
60. National Institute of Standards and Technology. Towards a standard for identifying and managing bias in artificial intelligence. <https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf>; Published March 2022. Accessed August 13, 2025.
61. National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf>; Published 2023. Accessed August 13, 2025.
62. World Health Organization. Regulatory considerations on artificial intelligence for health. <https://iris.who.int/bitstream/handle/10665/373421/9789240078871-eng.pdf?sequence=1>; Published 2023. Accessed August 13, 2025.
63. World Health Organization. Ethics and governance of artificial intelligence for health. Guidance on large multi-modal models. <https://iris.who.int/bitstream/handle/10665/375579/9789240084759-eng.pdf?sequence=1>; Published 2024. Accessed May 22, 2024.
64. Tripathi S, Alkhulaifat D, Doo FX, et al. Development, evaluation, and assessment of large language models (DEAL) checklist: a technical report. *NEJM AI*. 2025;2(6):1–6.
65. Ostmeier S, Xu J, Chen Z, et al. GREEN: Generative Radiology Report Evaluation and Error Notation. *arXiv*. published online May 6, 2024. <https://doi.org/10.48550/arXiv.2405.03595>.
66. Beam AL, Manrai AK, Ghassemi M. Challenges to the reproducibility of machine learning models in health care. *JAMA*. 2020;323(4):305–306.
67. Li S, Stenzel L, Eickhoff C, Bahrainian SA. Enhancing retrieval-augmented generation: a study of best practices. *arXiv*. published online January 13, 2025. <https://doi.org/10.48550/arXiv.2501.07391>.
68. Salemi A, Zamani H. *Evaluating retrieval quality in retrieval-augmented generation*. Washington, DC: Presented at: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2024.
69. 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;104646.
70. Xu J, Xiao Y, Wang WH, et al. Algorithmic fairness in computational medicine. *Ebiomedicine*. 2022;84:104250.
71. 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.
72. Zhao T, Wei M, Preston JS, Poon H. Automatic calibration and error correction for large language models via pareto optimal self-supervision. *arXiv*. published online March 3, 2024. <https://doi.org/10.48550/arXiv.2306.16564>.