

# The TRIPOD-LLM Statement: A Targeted Guideline For Reporting Large Language Models Use

Jack Gallifant<sup>1,2</sup>, Majid Afshar<sup>3\*</sup>, Saleem Ameen<sup>1,4,5\*</sup>, Yindalon Aphinyanaphongs<sup>6\*</sup>, Shan Chen<sup>7,8\*</sup>, Giovanni Cacciamani<sup>9,10\*</sup>, Dina Demner-Fushman<sup>11\*</sup>, Dmitriy Dligach<sup>12\*</sup>, Roxana Daneshjou<sup>13,14\*</sup>, Chrystinne Fernandes<sup>1\*</sup>, Lasse Hyldig Hansen<sup>1,15\*</sup>, Adam Landman<sup>16\*</sup>, Lisa Lehmann<sup>16\*</sup>, Liam G. McCoy<sup>17\*</sup>, Timothy Miller<sup>18\*</sup>, Amy Moreno<sup>19\*</sup>, Nikolaj Munch<sup>1,15\*</sup>, David Restrepo<sup>1,20\*</sup>, Guergana Savova<sup>18\*</sup>, Renato Umeton<sup>21\*</sup>, Judy Wawira Gichoya<sup>22\*</sup>, Gary S. Collins<sup>23,24</sup>, Karel G. M. Moons<sup>25,26</sup>, Leo A. Celi<sup>1,27,28</sup>, Danielle S. Bitterman<sup>7,8#</sup>

## Affiliations

1. Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA, USA.
2. Department of Critical Care, Guy's and St Thomas' NHS Foundation Trust, London, United Kingdom.
3. Department of Medicine, University of Wisconsin-Madison, Madison, WI 53705, United States
4. Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.
5. Tasmanian School of Medicine, College of Health and Medicine, University of Tasmania, Hobart, Australia.
6. Department of Population Health, NYU Grossman School of Medicine and Langone Health, New York, NY, USA
7. Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, MA, USA
8. Department of Radiation Oncology, Brigham and Women's Hospital/Dana-Farber Cancer Institute, Boston, MA, USA
9. USC Institute of Urology and Catherine and Joseph Aresty Department of Urology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA.
10. Artificial Intelligence Center, USC Institute of Urology, University of Southern California, Los Angeles, CA, USA.
11. National Library of Medicine, NIH, HHS, Bethesda, MD, USA
12. Department of Computer Science, Loyola University, Chicago, IL, United States
13. Department of Dermatology, Stanford School of Medicine, Redwood City, California, USA;
14. Department of Biomedical Data Science, Stanford School of Medicine, Redwood City, California, USA.

**NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.**

- 35 15. Cognitive Science, Aarhus University, Jens Chr. Skou 2, 8000 Aarhus, Denmark
- 36 16. Mass General Brigham, Boston, MA, United States
- 37 17. Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Alberta, Canada
- 38 18. Computational Health Informatics Program, Boston Children's Hospital, Harvard Medical School,  
39 Boston, MA, USA
- 40 19. Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center,  
41 Houston, Texas
- 42 20. Departamento de Telematica, Universidad del Cauca, Popayan, Cauca, Colombia.
- 43 21. Dana-Farber Cancer Institute, Boston, USA
- 44 22. Department of Radiology, Emory University School of Medicine, Atlanta, GA, USA
- 45 23. Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology, and  
46 Musculoskeletal Sciences, University of Oxford, Oxford OX3 7LD, United Kingdom
- 47 24. UK EQUATOR Centre, Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal  
48 Sciences, University of Oxford, Oxford OX3 7LD, United Kingdom
- 49 25. Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht University, Utrecht, The  
50 Netherlands.
- 51 26. Health Innovation Netherlands (HINL), The Netherlands.
- 52 27. Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center,  
53 Boston, MA, USA
- 54 28. Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

55

56 \* Equal contribution

57 # Corresponding Author

58 **COI**

59 DSB: Editorial, unrelated to this work: Associate Editor of Radiation Oncology, HemOnc.org (no financial  
60 compensation); Research funding, unrelated to this work: American Association for Cancer Research;  
61 Advisory and consulting, unrelated to this work: MercurialAI. DDF: Editorial, unrelated to this work:  
62 Associate Editor of JAMIA, Editorial Board of Scientific Data, Nature; Funding, unrelated to this work: the  
63 intramural research program at the U.S. National Library of Medicine, National Institutes of Health. JWG:  
64 Editorial, unrelated to this work: Editorial Board of Radiology: Artificial Intelligence, British Journal of  
65 Radiology AI journal and NEJM AI. All other authors declare no conflicts of interest.

66

67 **Abstract**

68 Large Language Models (LLMs) are rapidly being adopted in healthcare, necessitating standardized  
69 reporting guidelines. We present TRIPOD-LLM, an extension of the TRIPOD+AI statement, addressing the  
70 unique challenges of LLMs in biomedical applications. TRIPOD-LLM provides a comprehensive checklist  
71 of 19 main items and 50 subitems, covering key aspects from title to discussion. The guidelines introduce a  
72 modular format accommodating various LLM research designs and tasks, with 14 main items and 32  
73 subitems applicable across all categories. Developed through an expedited Delphi process and expert  
74 consensus, TRIPOD-LLM emphasizes transparency, human oversight, and task-specific performance  
75 reporting. We also introduce an interactive website (<https://tripod-llm.vercel.app/>) facilitating easy guideline  
76 completion and PDF generation for submission. As a living document, TRIPOD-LLM will evolve with the  
77 field, aiming to enhance the quality, reproducibility, and clinical applicability of LLM research in healthcare  
78 through comprehensive reporting.

79

80

## 81 **Introduction**

82 Healthcare's embrace of Large Language Models (LLMs) shows no signs of slowing down, with current  
83 and future deployment being considered in several domains across administrative and healthcare delivery  
84 use-cases, including in-basket draft generation, medical document summarization, question answering,  
85 information retrieval, medical diagnosis, treatment recommendations, patient education, and medical  
86 education.<sup>1-5</sup> The rapid advancements made in LLMs have stretched existing regulatory and governance  
87 structures to their limits, exposing a patchwork of solutions that do not fully encompass the nuances of these  
88 all-purpose models.<sup>6-8</sup> More broadly, with this speed, LLMs have posed a challenge to journal and peer-  
89 review publication timelines and challenged regulatory agencies to provide timely guidance. To maintain  
90 the speed, researchers publish pre-prints quickly and take an ad-hoc approach to reporting.

91 Reporting guidelines provide a scalable method for standardizing research, transparent reporting, and the  
92 peer review process. The TRIPOD (Transparent Reporting of a Multivariable Model for Individual  
93 Prognosis Or Diagnosis) initiative is a critical example that was first introduced in 2015 to establish  
94 minimum reporting standards for diagnostic and prognostic prediction model studies ([www.tripod-](http://www.tripod-statement.org)  
95 [statement.org](http://www.tripod-statement.org)).<sup>9</sup> TRIPOD is one of the core guidelines on the EQUATOR Network, which is an international  
96 effort that promotes transparent, accurate reporting of health research literature.<sup>10</sup> TRIPOD is widely  
97 endorsed and recommended by journals, and is often included in journal instructions to authors. TRIPOD  
98 has subsequently been updated to incorporate best practices for AI due to the significantly evolved machine  
99 learning landscape, resulting in TRIPOD+AI.<sup>11</sup> This is in addition to other guidelines that offer  
100 complementary guidance on AI development throughout the model lifecycle.<sup>12-14</sup> However, LLMs represent  
101 a distinct frontier within AI, introducing unique challenges and considerations not fully addressed by  
102 original TRIPOD guidelines or its newer extensions as we shift from classifier AI models to generative AI.  
103 Here, we report the TRIPOD+LLM statement, an extension of TRIPOD+AI<sup>11</sup>, developed to address these  
104 unmet needs and designed to be a living checklist in order to nimbly adapt to the rapidly evolving field. A  
105 completed example from a recent LLM research study is provided to guide future users (Supplementary  
106 Table 1).

107

### 108 *LLMs for biomedicine introduce unique complexities*

109 LLMs as generative AI models are autoregressive, meaning they are trained to predict the next word in a  
110 sequence given the words that preceded it. Yet, this foundational training has been shown to equip them  
111 with capabilities to perform a wide range of healthcare-related natural language processing (NLP) tasks from  
112 a single model. This adaptability is commonly achieved through supervised fine-tuning (SFT) or few-shot  
113 learning methods, which allow LLMs to handle new tasks with minimal examples.<sup>15,16</sup> Chatbot solutions  
114 (e.g., ChatGPT) use LLMs as their foundation, upon which two more components are added: question-  
115 answering (referred to as instruction tuning or supervised fine-tuning) and preference ranking (referred to  
116 as alignment). The unique methodological processes involved in LLMs and chatbots are not captured by  
117 current guidelines, such as the choice of hyperparameters used for SFT, the intricacies of prompting,  
118 variability in model predictions, methods in evaluating natural language outputs, and preference-based

119 learning strategies, which require specific guidance and significantly impact model reliability. In addition,  
120 the generalist and generative nature of LLMs require more detailed guidance than covered in prior  
121 guidelines. Because LLMs can be applied to a broad range of use cases for which they were not specifically  
122 trained for and were not necessarily represented in training data (e.g. disease prevalence typically captured  
123 in a task-specific model's training data for a given use case), they require unique task-specific guidance for  
124 robust reporting and downstream reliability and safety.

125 The selection of appropriate automated and human metrics by which to evaluate generative output remains  
126 an open question, and currently a wide range of methodologies are applied to capture various facets of  
127 performance. For tasks where the output is truly unstructured text and cannot be resolved to a structured  
128 label, evaluation is particularly complex. In these cases, most automated metrics prioritize overlap and  
129 similarity between input and output text, producing scores that may not accurately capture factual accuracy  
130 or relevance of the text produced, especially hallucinations or omissions.<sup>17-19</sup> These scores reflect the degree  
131 of structural and lexicographical resemblance which, though important, capture only a fraction of what  
132 constitutes a comprehensive evaluation of performance and safety. Human evaluation of text is subject to  
133 subjective interpretation, complicated by the ambiguity of language and uncertainties inherent to many  
134 clinical tasks. These challenges are heightened in medicine, where there is often no single correct answer  
135 and both aleatoric and epistemic uncertainty are common. Therefore, more details are needed to guide  
136 reporting of how performance is evaluated. In this paper, we use the term LLM to refer to both LLMs and  
137 chatbots. Table 1 highlights key categories of tasks applicable to the healthcare domain and provides notable  
138 definitions and examples of existing relevant work.

139 The novel complexities introduced by LLMs include concerns regarding hallucinations, omissions,  
140 reliability, explainability, reproducibility, privacy, and biases being propagated downstream, which can  
141 adversely affect clinical decision-making and patient care.<sup>20-26</sup> Furthermore, growing partnerships between  
142 EHR vendors, technology companies, and healthcare providers have led to deployment horizons that far  
143 outpace current regulatory timelines.<sup>8,27</sup> To safeguard LLM use and increase transparency, standardization  
144 in developing and reporting LLMs is essential to ensure consistency, reliability, and verifiability, akin to  
145 established clinical grade evaluation in other scientific domains.<sup>28-30</sup>

146

## 147 **Results**

### 148 *The TRIPOD-LLM Statement*

149 The TRIPOD-LLM comprises a checklist of items considered essential for good reporting of studies that are  
150 developing, tuning, prompt engineering, or evaluating an LLM (Table 2). Box 2 summarizes noteworthy  
151 additions and changes to TRIPOD-2015 and TRIPOD+AI. The TRIPOD-LLM checklist comprises 19 main  
152 items about the title (1 item), abstract (1 item), introduction (2 items), methods (8 items), open science  
153 practices (1 item), patient and public involvement (1 item), results (3 items), and discussion (2 items). These  
154 main items are further divided into 50 subitems. Of these, 14 main items and 32 subitems are applicable to  
155 all research designs and LLM tasks. The remaining 5 main items and 18 subitems are specific to particular  
156 research designs or LLM task categories. As discussed in the methods, the TRIPOD-LLM statement

157 introduces a modular format given the varied nature of LLM studies (Table 1), where some items are only  
158 relevant for specific research designs and LLM task categories.

159 A separate checklist for journal or conference abstracts of LLM-based studies is included, and the  
160 TRIPOD+AI for Abstracts statement<sup>18</sup> is revised (TRIPOD-LLM for Abstracts), reflecting new content and  
161 maintaining consistency with TRIPOD-LLM (Table 3).

162 The recommendations contained within TRIPOD-LLM are for completely and transparently reporting on  
163 how LLM-based research was conducted; TRIPOD-LLM does not prescribe how to develop or evaluate  
164 LLMs specifically. The checklist is not a quality appraisal tool. Similarly, CANGARU<sup>31</sup> and CHART<sup>32</sup> are  
165 complementary guidelines that relate to Generative AI more broadly and Chatbots specifically.

166 In addition to the TRIPOD website ([www.tripod-statement.org](http://www.tripod-statement.org)) an accompanying interactive website was  
167 developed (<https://tripod-llm.vercel.app/>) to present the required questions based on research design(s) and  
168 task(s) for ease of completion. This site can be used to render a final PDF suitable for submission. Fillable  
169 templates for the TRIPOD-LLM checklist can also be found in the supplementary material or downloaded  
170 from [www.tripod-statement.org](http://www.tripod-statement.org). News, announcements, and information relating to TRIPOD-LLM and the  
171 release of subsequent statements can be found on the TRIPOD-LLM website, TRIPOD website  
172 ([www.tripod-statement.org](http://www.tripod-statement.org)), and on social media accounts such as X (formerly known as Twitter)  
173 @TRIPODStatement.

174 An example of a completed TRIPOD-LLM checklist for a previously published study reporting the pre-  
175 training, fine-tuning, retrospective evaluation, and clinical deployment of an LLM for clinical and  
176 operational hospital tasks is presented in Table 5. The design categories relevant to this work are *de novo*  
177 LLM development, LLM evaluation, and LLM evaluation in healthcare settings. Task categories relevant to  
178 this work are classification and outcome forecasting.

179

### 180 TRIPOD-LLM Statement as a Living Document

181 Given the rapid pace of the field and the timeline for interaction with healthcare workers and patients, the  
182 decision was made to create an accelerated TRIPOD-LLM statement to provide timely guidance for LLM  
183 use in (bio)medical and other healthcare applications. This guidance has been designed as part of a living  
184 document hosted on an interactive website to facilitate agile versioning, refinement from user testing,  
185 updates as the field evolves, and regular meetings to intake and evaluate new standards. Thus, as the  
186 reporting recommendations are anticipated to evolve; users are directed to the most current version of the  
187 guidelines at <https://tripod-llm.vercel.app/>.

188 Our approach to the living TRIPOD-LLM statement is informed by processes established in developing  
189 living systematic reviews<sup>33,33,34</sup> and clinical practice guidelines,<sup>35,36</sup> which have been adopted to address a  
190 similar need to provide updated, timely recommendations based on evolving evidence. Public comments on  
191 the statement will be collected from the community via multiple avenues to enhance accessibility: a project-  
192 specific GitHub repository, the TRIPOD-LLM website, and the main TRIPOD website ([https://www.tripod-  
193 statement.org/](https://www.tripod-statement.org/)). We encourage input both on usability, such as language that may be ambiguous or

194 redundant, and on the content of the guidelines themselves. As a few examples, users may suggest a change  
195 to an item to make it more feasible in practice, recommend a new item be added, recommend adding or  
196 removing items assigned to a given research design of LLM task module, or recommend changes to the  
197 research design or LLM task module categories.

198

199 An expert panel will convene every three months to discuss updates. Before the meeting, members will  
200 review the intercurrent literature to inform any updates. The units for update will be checklist items, research  
201 design categories, and LLM task categories delineated in the statement. At the meeting, the panel will discuss  
202 the current statement and suggest revisions considering public comments, literature review, and subject  
203 matter expertise. The steering committee will revise the statement based on this discussion, and this will be  
204 circulated to the expert panel for final review and approval. Review can result in the following action for  
205 each component of the TRIPOD-LLM statement (adapted from Mikati et al., 2019<sup>33</sup>) - items, research  
206 designs, and LLM tasks:

- 207 1. No modification
- 208 2. Modification of substantive content (small, editorial revisions such as re-wording for clarity and and  
209 correcting types will not be considered a modification)
- 210 3. Merging of one or more components together (merging will only take place within a component  
211 type)
- 212 4. Splitting one component into two or more components (splitting will only take place within a  
213 component type)
- 214 5. Retiring the component from the statement

215 Release of a new version of the statement will be disseminated to the community through postings on the  
216 TRIPOD-LLM website, the main TRIPOD website (<https://www.tripod-statement.org/>), the EQUATOR  
217 Network website (<https://www.equator-network.org/reporting-guidelines/>), and postings on social media  
218 accounts such as X (formerly known as Twitter) @TRIPODStatement. Emails will be sent to journal editors  
219 to inform them of the update and ensure that author instructions refer to the most current versioning. The  
220 detailed transcripts of the discussion are available in supplementary materials for full transparency.

221

222 At each review meeting, the membership of the expert panel will be reviewed for expertise, diversity, and  
223 representation, and new members will be solicited if and when gaps are identified. Expert panel members  
224 will also have the authority to trigger an ad hoc review of the guidelines to accommodate major, unexpected  
225 changes in the field that warrant more urgent discussion.

226

## 227 **Discussion**

228 TRIPOD-LLM has been developed to guide researchers, journals, healthcare professionals, LLM  
229 developers (commercially and non-commercially), and healthcare institutions in the rapidly evolving field  
230 of biomedical and healthcare LLMs. It represents minimum reporting recommendations for studies

231 describing LLMs' development, tuning, or evaluation. Reporting TRIPOD-LLM items will help users  
232 understand and appraise the quality of LLM study methods, increase transparency around study findings,  
233 reduce overinterpretation of study findings, facilitate replication and reproducibility, and aid in  
234 implementing the LLM.

235 **Transparency throughout the model lifecycle** has been emphasized significantly in the guidelines.  
236 Detailed documentation is emphasized at each stage of an LLM's life cycle;<sup>37</sup> for example, during the  
237 development and fine-tuning phases, there is an emphasis on disclosing the origins and processing of training  
238 data. Moreover, the LLM version and specifics of any fine-tuning or alignment modeling processes on top  
239 of existing foundation models must be transparently reported to enable fair comparisons of LLMs. This  
240 includes specifying the cut-off dates for when training data were collected to clarify the temporal relevance  
241 of training datasets and potential for data leakage or contamination during evaluation. In addition, the model  
242 version date and if the model was frozen or remained dynamic during the data collection phase from  
243 generated output should be documented. Transparency regarding data is essential because LLMs are  
244 typically trained on multiple public large-scale datasets and thus inherently risk incorporating existing  
245 societal biases and inequities in stigmatizing language use as well as statistical risk allocation in disparate  
246 groups, necessitating a comprehensive transparency approach to training data sources and content to  
247 understand potential biases.<sup>21,22,38-41</sup>

248 **Human insight and oversight** are critical components of the TRIPOD-LLM statement, reflecting an  
249 emphasis on components eventually critical for the responsible deployment of LLMs (though deployment  
250 reliability and observability are outside the scope of this paper).<sup>42-44</sup> The guidelines include requirements for  
251 increased reporting of the expected deployment context and specifying the levels of autonomy assigned to  
252 the LLM, if applicable. Furthermore, there is a focus on the quality control processes employed in dataset  
253 development and evaluation, such as qualifications of human assessors, requirement for dual annotation,  
254 and specific details on instructions provided to assessors to ensure that nuances of text evaluation are  
255 captured, thus facilitating reliable assessments of safety and performance.

256 **Prompting and task-specific performance** are key additions necessitated by the unique characteristics of  
257 LLMs. The variability in prompt engineering approaches can significantly influence LLM performance,  
258 potentially skewing benchmark comparisons and real-world applicability.<sup>45,46</sup> Reports, where relevant, must  
259 include comprehensive descriptions of data sources used for developing prompts, LLM model name and  
260 versions, any preprocessing undertaken, and methods employed in prompt engineering. This ensures that  
261 prompts are effectively designed to elicit stable and reproducible performance from LLMs. Additionally,  
262 the guidelines call for clear reporting on evaluation settings, including instructions and interfaces used and  
263 characteristics of populations involved in evaluations. This is intended to ensure that LLM performance is  
264 assessed under conditions that closely mimic real-world applications, providing a reliable measure of its  
265 practical utility.

266 We anticipate that key users and beneficiaries of TRIPOD-LLM will be (1) academic and industry  
267 researchers authoring papers, (2) journal editors and peer reviewers evaluating research papers, and (3) other  
268 stakeholders (e.g., the research community in general, academic institutions, policy-makers, funders,



269 regulators, patients, study participants, industry, and the broader public) who will benefit from increased  
270 transparency and quality of LLM research. We encourage editors, publishers, and the industry more broadly  
271 to support adherence to TRIPOD-LLM by referring to a link within the journal's instructions to authors,  
272 enforcing its use during the submission and peer review process, and making adherence to the  
273 recommendations an expectation. We also encourage funders to require applications for LLM studies to  
274 include a plan to report their model according to the TRIPOD-LLM recommendations, thereby minimizing  
275 research waste and ensuring value for money.

276 Of note, this guideline was developed with text-only LLMs in mind; however, advances in multi-modal  
277 models incorporating LLMs, such as vision-language models,<sup>47</sup> are now rapidly emerging - illustrating the  
278 need for rapid, nimble approaches for reporting guidelines. Many of the reporting considerations will be  
279 shared between text-only LLMs and these multi-modal models. For example, for vision-language models,  
280 both text and image preprocessing should be reported. However, unique considerations may arise that merit  
281 discussion in future versioning of TRIPOD-LLM or related guidelines. For example, studies reporting the  
282 development of LLMs that use imaging data should report details of image acquisition. In the interim, we  
283 suggest that studies reporting the development and/or evaluation of a method that includes an LLM as a  
284 primary component use the TRIPOD-LLM statement, although we acknowledge this may be subject to  
285 interpretation. We advise that users keep in mind the goals of reproducibility, understandability, and  
286 transparency to take a common-sense approach to deciding on the appropriate reporting guideline, and to  
287 interpreting the relevant components of TRIPOD-LLM statement items to report multimodal LLMs. Users  
288 may also refer to methodological guides from multiple AI fields, such as radiomics,<sup>48,49</sup> to inform their  
289 reporting.

290 **The role of assurance labs** such as the Coalition for Health AI (CHAI)<sup>50</sup> and Epic AI Labs,<sup>51,52</sup> or internal  
291 validation standards are expected to be of importance in the generation, verification, certification, and  
292 maintenance of model cards for clinical AI. It is our opinion that the TRIPOD-LLM standard can and should  
293 inform such assurance labs as they develop approaches to assure LLMs in ways that meet the required  
294 regulatory bar for AI (e.g., the Biden-Harris Administration “Executive Order on Safe, Secure, and  
295 Trustworthy Artificial Intelligence”, the United States AI Safety Institute,<sup>53</sup> the United States Office of the  
296 National Coordinator for Health Information Technology HTI-1 Final Rule,<sup>54</sup> and the European Union AI  
297 Act<sup>55</sup>) and build confidence in patients, clinician, and other stakeholders about the utility and trustworthiness  
298 of clinical AI.

299 It must also be emphasized that LLM evaluation and validation requires specialized expertise and resources.  
300 To ensure equitable and safe deployment, investments into LLM development should be balanced by  
301 investments into infrastructure that enables robust validation beyond large academic settings. Moreover, this  
302 checklist should be seen as part of a continuous process for evaluating LLMs due to the temporal and  
303 geographic specific contexts these models inherit, which can impact the generalisability of performance and  
304 fairness across sites or at the same site over time. These shifts can be even more unpredictable than  
305 traditional ML models due to their user-dependent nature, and thus, significant effort must be placed on  
306 understanding trends and heterogeneity of effects instead of single-point estimates that proclaim universal  
307 validation.

308

309 Conclusion

310 TRIPOD-LLM aims to assist authors in the complete reporting of their study and help LLM developers,  
311 researchers, peer reviewers, editors, policymakers, end-users (e.g., healthcare professionals), and patients  
312 understand data, methods, findings, and conclusions of LLM-driven research. Adhering to the TRIPOD-  
313 LLM reporting recommendations may promote the best and most efficient use of research time, effort, and  
314 money, enhancing the value of LLM research to maximize positive impact.

315

316 **Methods**

317 The TRIPOD-LLM statement was formulated to guide the reporting of studies that develop, tune, or evaluate  
318 LLMs for any healthcare application or context and was crafted following pathways utilized in creating the  
319 other TRIPOD statements. An expedited Delphi process was implemented given the need for timely  
320 reporting guidelines in this field, and the living statement approach. An accompanying glossary (Box 1)  
321 defines essential terms relevant to the TRIPOD-LLM statement.

322 A steering group, including DSB, JG, LAC, GSC, and KGMM, was established to direct the guideline  
323 development process. They were joined by an expert panel, including SC, CF, DR, GS, TM, DFD, RU,  
324 LHH, YA, JWG, LGM, NM, and RD, and were chosen based on their diverse expertise and experience in  
325 natural language processing, artificial intelligence, and medical informatics. This guideline was registered  
326 on May 2, 2024, with the EQUATOR Network as a reporting guideline under development ([www.equator-](http://www.equator-network.org)  
327 [network.org](http://www.equator-network.org)).

328 Ethics

329 This study received an exemption from the MIT COUHES IRB review board on March 26, 2024 (Exempt  
330 ID: E-5705). Delphi survey participants provided electronic informed consent before completing the survey.

331

332 Candidate item list generation

333 The TRIPOD-2015 and TRIPOD+AI guidelines ([www.tripod-statement.org](http://www.tripod-statement.org)) and associated literature on  
334 reporting guidelines for LLMs were used to inform the initial candidate item list drafted by DSB and  
335 JG.<sup>9,11,28,32</sup> The steering group and expert panel expanded this list through additional literature review,  
336 ultimately standardizing it to 64 unique items across the following sections: title, abstract, introduction,  
337 methods, results, discussion, and others.

338

339 Panelist recruitment

340 Delphi participants were identified by the steering committee from authors of relevant publications and  
341 through personal recommendations, including experts recommended by other Delphi participants. The  
342 steering group identified participants representing geographical and disciplinary diversity, including key  
343 stakeholder groups, e.g., researchers (statisticians, data scientists, epidemiologists, machine learners,

344 clinicians, and ethicists), healthcare professionals, journal editors, funders, policymakers, healthcare  
345 regulators, and patient advocate groups. No minimum sample size was placed on the number of participants.  
346 A steering group member checked the expertise or experience of each identified person. Participants were  
347 then invited to complete a survey via e-mail. Delphi participants did not receive any financial incentive or  
348 gift to participate.

### 349 Delphi process

351 The survey was designed to allow individual responses, in English and delivered electronically using Google  
352 Forms. All responses were anonymous; no emails or identifying information was collected from participants.  
353 Participants were asked to rate each item as ‘can be omitted,’ ‘possibly include,’ ‘desirable for inclusion,’  
354 or ‘essential for inclusion’ as has been conducted in previous TRIPOD guidelines.<sup>9</sup> Participants were also  
355 invited to comment on any item and suggest new items. DSB and JG collated and analyzed the free text  
356 responses then used the themes generated to inform item rephrasing, merging, or suggesting new items. All  
357 steering group members were invited to participate in the Delphi surveys.

### 358 Round 1 participants

360 The first round was conducted between 01 March 2024 and 23 April 2024, where the participation link was  
361 sent to 56 people. Of the 56 invited, 26 completed the survey. Survey participants came from 9 countries,  
362 of which 14 were from North America, five from Europe, two from Asia, one from South America, and one  
363 from Australasia. Three participants did not provide this information. Participants reported their primary  
364 fields of work and could select more than one field. 20/26 (77 %) had a primary field in AI, Machine  
365 Learning, Clinical Informatics, or NLP, and 14/26 selected healthcare.

### 366 Consensus Meeting

368 An online consensus meeting was held on the 22nd and 24th of April, chaired by DSB and JG via Zoom.  
369 All steering committee and expert panel members were invited to attend. Recordings and notes were sent  
370 immediately after the meetings to enable asynchronous contribution for those who did not attend. The  
371 responses to each question were examined in turn, as well as all free-text comments. Items with <50%  
372 ‘Essential to Include’ were highlighted and deliberated for the importance of inclusion. Agreement by  
373 consensus without needing a third party was reached in all cases. To arrive at a consensus, the item was  
374 discussed until no panel had additional comments or disagreements with the final disposition of the item.

375 Due to the vast number of applications being developed using LLMs, a modular approach was used to group  
376 included items under additional subcategories under the ‘Research Design’ and ‘LLM Task’ headers. This  
377 approach was agreed upon during the meeting, and the steering committee approved the final groupings.

378 To adequately address the variety of studies and uses regarding LLMs, ranging across stages of  
379 development, tuning, evaluation, and implementation, as well as across healthcare tasks, items are  
380 categorized according to (1) research design and (2) LLM task. The research design categories are: *de novo*

381 LLM development, LLM methods such as fine-tuning, prompt-engineering techniques and architecture  
382 modifications, intrinsic LLM evaluation, and LLM evaluation in dedicated healthcare settings and tasks.  
383 The LLM task categories are lower-level text processing (e.g., part-of-speech tagging, relation extraction,  
384 named entity recognition), classification (e.g., diagnosis), long-form question-answering, conversational  
385 agent, documentation generation, summarization/simplification, machine translation, and outcome  
386 forecasting (e.g., prognosis). Items may apply to several design and task categories, and studies may include  
387 more than one design and task. Items applicable to every type and task covered in the study should be  
388 reported. Definitions and examples of each design and task category are provided in Table 1. We  
389 acknowledge that these categorizations are imperfect and overlap may exist across designs and tasks.

390

## 391 **Acknowledgments**

392 JG is funded by the NIH-USA U54 TW012043-01 and NIH-USA OT2OD032701.

393 SC was supported by NIH-USA U54CA274516-01A1.

394 GSC was supported by Cancer Research UK (programme grant: C49297/A27294), and by EPSRC grant for  
395 ‘Artificial intelligence innovation to accelerate health research’ (number: EP/Y018516/1, and is a National  
396 Institute for Health and Care Research (NIHR) Senior Investigator. The views expressed in this article are  
397 those of the author and not necessarily those of the NIHR, or the Department of Health and Social Care.

398 Yin Aphinyanaphongs was partially supported by NIH 3UL1TR001445-05 and National Science  
399 Foundation award #1928614 & #2129076.

400 LAC is funded by NIH-USA U54 TW012043-01, NIH-USA OT2OD032701, and NIH-USA  
401 R01EB017205.

402 DSB was supported by NIH-USA U54CA274516-01A1 and NIH-USA R01CA294033-01.

403 Dr. Gichoya is a 2022 Robert Wood Johnson Foundation Harold Amos Medical Faculty Development  
404 Program and declares support from RSNA Health Disparities grant (#EIHD2204), Lacuna Fund (#67),  
405 Gordon and Betty Moore Foundation, NIH (NIBIB) MIDRC grant under contracts 75N92020C00008 and  
406 75N92020C00021, and NHLBI Award Number R01HL167811.

## 407 **References**

- 408 1. Chen, Z. *et al.* MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. Preprint at  
409 <https://doi.org/10.48550/arXiv.2311.16079> (2023).
- 410 2. OpenAI. GPT-4 Technical Report. Preprint at <https://doi.org/10.48550/arXiv.2303.08774> (2023).
- 411 3. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
- 412 4. Tai-Seale, M. *et al.* AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic  
413 Communication. *JAMA Netw. Open* **7**, e246565 (2024).
- 414 5. Tierney, A. A. *et al.* Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical  
415 Documentation. *NEJM Catal.* **5**, CAT.23.0404 (2024).
- 416 6. Jiang, L. Y. *et al.* Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–  
417 362 (2023).
- 418 7. Regulating advanced artificial agents | Science. <https://www.science.org/doi/10.1126/science.adl0625>.
- 419 8. Meskó, B. & Topol, E. J. The imperative for regulatory oversight of large language models (or generative AI)  
420 in healthcare. *Npj Digit. Med.* **6**, 1–6 (2023).
- 421 9. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable  
422 prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* **350**, g7594  
423 (2015).
- 424 10. Reporting guidelines | EQUATOR Network. <https://www.equator-network.org/reporting-guidelines/>.
- 425 11. Collins, G. S. *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use  
426 regression or machine learning methods. *The BMJ* **385**, e078378 (2024).
- 427 12. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J. & Denniston, A. K. Reporting guidelines for clinical trial  
428 reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* **26**, 1364–  
429 1374 (2020).
- 430 13. Vasey, B. *et al.* Reporting guideline for the early-stage clinical evaluation of decision support systems driven  
431 by artificial intelligence: DECIDE-AI. *Nat. Med.* **28**, 924–933 (2022).
- 432 14. Norgeot, B. *et al.* Minimum information about clinical artificial intelligence modeling: the MI-CLAIM  
433 checklist. *Nat. Med.* **26**, 1320–1324 (2020).
- 434 15. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y. & Sontag, D. Large Language Models are Few-Shot Clinical  
435 Information Extractors. Preprint at <https://doi.org/10.48550/arXiv.2205.12689> (2022).
- 436 16. Liu, X. *et al.* Large Language Models are Few-Shot Health Learners. Preprint at

- 437 <https://doi.org/10.48550/arXiv.2305.15525> (2023).
- 438 17. Salazar, J., Liang, D., Nguyen, T. Q. & Kirchhoff, K. Masked Language Model Scoring. in *Proceedings of the*  
439 *58th Annual Meeting of the Association for Computational Linguistics* 2699–2712 (2020).  
440 doi:10.18653/v1/2020.acl-main.240.
- 441 18. Wang, A. *et al.* GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding.  
442 Preprint at <https://doi.org/10.48550/arXiv.1804.07461> (2019).
- 443 19. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a Method for Automatic Evaluation of Machine  
444 Translation. in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (eds.  
445 Isabelle, P., Charniak, E. & Lin, D.) 311–318 (Association for Computational Linguistics, Philadelphia,  
446 Pennsylvania, USA, 2002). doi:10.3115/1073083.1073135.
- 447 20. Goodman, K. E., Yi, P. H. & Morgan, D. J. AI-Generated Clinical Summaries Require More Than Accuracy.  
448 *JAMA* (2024) doi:10.1001/jama.2024.0555.
- 449 21. Zack, T. *et al.* Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model  
450 evaluation study. *Lancet Digit. Health* **6**, e12–e22 (2024).
- 451 22. Gallifant, J. *et al.* Peer review of GPT-4 technical report and systems card. *PLOS Digit. Health* **3**, 1–15 (2024).
- 452 23. Wornow, M. *et al.* The shaky foundations of large language models and foundation models for electronic  
453 health records. *Npj Digit. Med.* **6**, 1–10 (2023).
- 454 24. Chen, S. *et al.* The effect of using a large language model to respond to patient messages. *Lancet Digit. Health*  
455 **0**, (2024).
- 456 25. Chang, C. T. *et al.* Red Teaming Large Language Models in Medicine: Real-World Insights on Model  
457 Behavior. 2024.04.05.24305411 Preprint at <https://doi.org/10.1101/2024.04.05.24305411> (2024).
- 458 26. Gallifant, J. *et al.* Language Models are Surprisingly Fragile to Drug Names in Biomedical Benchmarks.  
459 Preprint at <https://doi.org/10.48550/arXiv.2406.12066> (2024).
- 460 27. Blogs, M. C. Microsoft and Epic expand AI collaboration to accelerate generative AI’s impact in healthcare,  
461 addressing the industry’s most pressing needs. *The Official Microsoft Blog*  
462 [https://blogs.microsoft.com/blog/2023/08/22/microsoft-and-epic-expand-ai-collaboration-to-accelerate-](https://blogs.microsoft.com/blog/2023/08/22/microsoft-and-epic-expand-ai-collaboration-to-accelerate-generative-ais-impact-in-healthcare-addressing-the-industrys-most-pressing-needs/)  
463 [generative-ais-impact-in-healthcare-addressing-the-industrys-most-pressing-needs/](https://blogs.microsoft.com/blog/2023/08/22/microsoft-and-epic-expand-ai-collaboration-to-accelerate-generative-ais-impact-in-healthcare-addressing-the-industrys-most-pressing-needs/) (2023).
- 464 28. Moreno, A. C. & Bitterman, D. S. Toward Clinical-Grade Evaluation of Large Language Models. *Int. J.*  
465 *Radiat. Oncol. Biol. Phys.* **118**, 916–920 (2024).
- 466 29. Spann, M. Welch Medical Library Guides: Evidence Based Medicine: Evidence Grading & Reporting.

- 467 [https://browse.welch.jhmi.edu/EBM/EBM\\_EvidenceGrading](https://browse.welch.jhmi.edu/EBM/EBM_EvidenceGrading).
- 468 30. Guyatt, G. H. *et al.* What is “quality of evidence” and why is it important to clinicians? *BMJ* **336**, 995–998  
469 (2008).
- 470 31. Cacciamani, G. E., Collins, G. S. & Gill, I. S. ChatGPT: standard reporting guidelines for responsible use.  
471 *Nature* **618**, 238–238 (2023).
- 472 32. Huo, B. *et al.* Reporting standards for the use of large language model-linked chatbots for health advice. *Nat.*  
473 *Med.* **29**, 2988–2988 (2023).
- 474 33. El Mikati, I. K. *et al.* A Framework for the Development of Living Practice Guidelines in Health Care. *Ann.*  
475 *Intern. Med.* **175**, 1154–1160 (2022).
- 476 34. Living systematic reviews | Cochrane Community. [https://community.cochrane.org/review-](https://community.cochrane.org/review-development/resources/living-systematic-reviews)  
477 [development/resources/living-systematic-reviews](https://community.cochrane.org/review-development/resources/living-systematic-reviews).
- 478 35. Akl, E. A. *et al.* Living systematic reviews: 4. Living guideline recommendations. *J. Clin. Epidemiol.* **91**, 47–  
479 53 (2017).
- 480 36. Fraile Navarro, D. *et al.* Methods for living guidelines: early guidance based on practical experience. Paper 5:  
481 decisions on methods for evidence synthesis and recommendation development for living guidelines. *J. Clin.*  
482 *Epidemiol.* **155**, 118–128 (2023).
- 483 37. Bedoya, A. D. *et al.* A framework for the oversight and local deployment of safe and high-quality prediction  
484 models. *J. Am. Med. Inform. Assoc. JAMIA* **29**, 1631–1636 (2022).
- 485 38. Omiye, J. A., Gui, H., Rezaei, S. J., Zou, J. & Daneshjou, R. Large Language Models in Medicine: The  
486 Potentials and Pitfalls. *Ann. Intern. Med.* (2024) doi:10.7326/M23-2772.
- 487 39. Chen, S. *et al.* Cross-Care: Assessing the Healthcare Implications of Pre-training Data on Language Model  
488 Bias. Preprint at <https://doi.org/10.48550/arXiv.2405.05506> (2024).
- 489 40. Hansen, L. H. *et al.* Seeds of Stereotypes: A Large-Scale Textual Analysis of Race and Gender Associations  
490 with Diseases in Online Sources. Preprint at <https://doi.org/10.48550/arXiv.2405.05049> (2024).
- 491 41. Biderman, S. *et al.* Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling.  
492 Preprint at <https://doi.org/10.48550/arXiv.2304.01373> (2023).
- 493 42. Bowman, S. R. *et al.* Measuring Progress on Scalable Oversight for Large Language Models. (2022)  
494 doi:10.48550/ARXIV.2211.03540.
- 495 43. McAleese, N. *et al.* LLM Critics Help Catch LLM Bugs. Preprint at  
496 <https://doi.org/10.48550/arXiv.2407.00215> (2024).

- 497 44. Burns, C. *et al.* Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. Preprint  
498 at <https://doi.org/10.48550/arXiv.2312.09390> (2023).
- 499 45. Chen, S. *et al.* Evaluating the ChatGPT family of models for biomedical reasoning and classification. *J. Am.*  
500 *Med. Inform. Assoc.* ocad256 (2024) doi:10.1093/jamia/ocad256.
- 501 46. Chen, S. *et al.* Use of Artificial Intelligence Chatbots for Cancer Treatment Information. *JAMA Oncol.* **9**,  
502 1459–1462 (2023).
- 503 47. Lu, M. Y. *et al.* A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863–874  
504 (2024).
- 505 48. Kocak, B. *et al.* METHodological RadiomICs Score (METRICS): a quality scoring tool for radiomics research  
506 endorsed by EuSoMII. *Insights Imaging* **15**, 8 (2024).
- 507 49. Lambin, P. *et al.* Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin.*  
508 *Oncol.* **14**, 749–762 (2017).
- 509 50. Shah, N. H. *et al.* A Nationwide Network of Health AI Assurance Laboratories. *JAMA* **331**, 245–249 (2024).
- 510 51. Epic releases AI validation suite. [https://www.beckershospitalreview.com/ehrs/epic-releases-ai-validation-](https://www.beckershospitalreview.com/ehrs/epic-releases-ai-validation-suite.html)  
511 [suite.html](https://www.beckershospitalreview.com/ehrs/epic-releases-ai-validation-suite.html).
- 512 52. epic-open-source/seismometer: AI model evaluation with a focus on healthcare. [https://github.com/epic-open-](https://github.com/epic-open-source/seismometer)  
513 [source/seismometer](https://github.com/epic-open-source/seismometer).
- 514 53. U.S. Artificial Intelligence Safety Institute. *NIST* (2023).
- 515 54. Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and  
516 Information Sharing. *Federal Register* [https://www.federalregister.gov/documents/2024/01/09/2023-](https://www.federalregister.gov/documents/2024/01/09/2023-28857/health-data-technology-and-interoperability-certification-program-updates-algorithm-transparency-and)  
517 [28857/health-data-technology-and-interoperability-certification-program-updates-algorithm-transparency-and](https://www.federalregister.gov/documents/2024/01/09/2023-28857/health-data-technology-and-interoperability-certification-program-updates-algorithm-transparency-and)  
518 (2024).
- 519 55. EU Artificial Intelligence Act | Up-to-date developments and analyses of the EU AI Act.  
520 <https://artificialintelligenceact.eu/>.



## The TRIPOD-LLM Statement: A Targeted Guideline For Reporting Large Language Models Use

<b>Box 1: Glossary of terms (in alphabetical order)</b>	
<b>Term</b>	<b>Definition</b>
Application Programming Interface (API)	A set of rules and protocols for building and interacting with software applications, allowing different software entities to communicate seamlessly.
Artificial Intelligence (AI) Models	Computational systems designed to simulate human intelligence tasks such as problem-solving, decision-making, and language processing.
Attention Mechanism	A component in neural networks that allows the model to focus on different parts of the input when producing each part of the output, crucial for handling long-range dependencies in sequence data.
Bias	A deviation in model output caused by systematic errors in the data or algorithmic prejudice, potentially leading to skewed or unfair outcomes in clinical predictions, risk assessments, and treatment recommendations.
Chain-of-Thought Prompting	A prompting technique that encourages the model to break down complex reasoning tasks into step-by-step thought processes, often improving performance on logical and mathematical tasks.
Clinical question-answering	The application of question-answering systems in clinical settings, providing answers to medically related questions based on large volumes of healthcare data.
Data Leakage	The use of test data during model training and/or fine-tuning, resulting in a model that will perform at a greater level than if it had been fully tested on unseen data.
Decoder	A component of a model that converts vectorized input data and converts it back into a text sequence.
Autoregressive model	A type of Transformer-based model that models predict the next component in a sequence, for example the next word in a sentence, based on the preceding sequence. Current state-of-the-art LLMs, including generative pre-trained transformers, are autoregressive models.
Domain Knowledge	Specialized knowledge or expertise in a particular field or subject area, including its concepts, principles, and applications.
Embedding	A representation of text in a high-dimensional vector space where similar words have similar representation, capturing semantic meaning. See also Vector.
Encoder	A component of a model that processes the input data, transforming it into a vectorized format or representation that the model can understand.
Encoder-Decoder	A model architecture framework combining an encoder and decoder to transform input data into an output.
Externalization	The process of converting tacit knowledge into explicit knowledge, making it accessible through various forms like documents or databases.
Few-Shot Learning	A method where the model learns to perform a task proficiently with a very small number of exemplars. In some cases, the number of exemplars are specified in place of “few”, e.g. (one-shot learning).
Fine-tuning	A process where a pre-trained model is further trained (or fine-tuned) on a smaller, domain-specific dataset to specialize its knowledge for specific tasks.
Generative Pre-trained Transformer (GPT)	A family of autoregressive Transformer-based models for natural language understanding and generation. These models are pre-trained to predict the next word in a sentence.
Hallucination	A phenomenon where a language model generates text that is unrelated or loosely related to the input data, often manifesting as fabrications or inaccuracies.
In Context Learning	The ability of a model to learn a new task from examples provided within this prompt at inference time.
Instruction Tuning	A fine-tuning approach where models are trained on datasets consisting of natural language instructions and their corresponding desired outputs, improving the model's ability to follow diverse instructions.
Intended Use	How the model is intended to be used, including the intended population, task, end-users, and level of human supervision.
Natural Language Processing (NLP) Models	AI models designed to process, understand, and generate human language, facilitating interactions between computers and humans.
Prompt	The query or instruction that is input into an LLM to elicit a response.
Reinforcement learning with human feedback	A machine learning technique commonly used in LLM development that trains a model to optimize its output according to humans’ preferences by providing rewards in response to actions.
Prompt engineering	A process that guides the models to generate desired outputs. Examples of prompt engineering include prompt iterations, prompting with examples, and chain-of-thought prompting.
Prompt Injection	A security vulnerability in LLM applications where an attacker attempts to manipulate the model's behavior by inserting malicious content into the input prompt.
Retrieval-Augmented Generation (RAG)	A technique that combines information retrieval from an external knowledge base with text generation, allowing LLMs to access and incorporate up-to-date or domain-specific information.
Simplification	The process of converting a text into a simplified version, commonly used in biomedicine to describe the conversion of technical language into patient-friendly, plain language versions.

Summarization	The process of condensing a larger text document into a shorter version, preserving key information and the overall meaning.
Temperature	A parameter that controls the randomness of predictions by scaling the logits before applying softmax, affecting the diversity of generated text.
Text annotation	The process of labeling a text dataset with the goal output label. The label is defined by the given task. Annotated datasets may be used to develop models and/or serve as benchmarks for model evaluation.
Tokenization	The process of converting text into smaller units, such as words or phrases, to facilitate their processing in NLP tasks.
Transformer	A commonly used neural network architecture that has significantly advanced the field of NLP. Unlike its predecessors, the transformer relies on self-attention mechanisms to process sequences of data in parallel, improving efficiency and the ability to capture complex dependencies within the text.
Vector	A numerical representation of data in the form that an AI model can process. In the context of LLMs, text data is represented as a type of vector known as contextual embeddings, meaning that the vector for each token (word piece) is influenced by the words surrounding it. For example, the word “gray” has different vectors when used in “gray matter”, “54 gray”, and “Gray’s anatomy”.
Zero-Shot Learning	The ability of a model to correctly perform tasks it has never explicitly been trained to do, based on its understanding and generalization capabilities.
The definitions and descriptions given relate to the specific context of TRIPOD-LLM and the use of the terms in the guideline. They are not necessarily applicable to other areas of research.	

## The TRIPOD-LLM Statement: A Targeted Guideline For Reporting Large Language Models Use

### Box 2: TRIPOD-LLM noteworthy changes and additions to TRIPOD-2015 and TRIPOD+AI

- **New Checklist for Reporting on LLMs:** A dedicated checklist has been developed to address the unique aspects of reporting large language models (LLMs), reflecting their distinct characteristics and the specific methodologies they employ compared to other AI and prediction models.
- **Living Guideline:** The checklist is designed as a living document, which will be updated on a regular basis based on review of the literature and input from the community. This approach was taken due to the rapid developments in the field, enabling agile versioning, refinement from user testing, and timely updates as the field advances.
- **Task-Specific Guidance:** The checklist includes a new section that provides task-specific guidance designed to address the particular challenges and needs associated with different LLM applications in healthcare. This addition ensures that reporting is tailored and relevant to the specific functions and objectives of the LLM under study.
- **Enhanced Emphasis on Transparency and Fairness:** The new guidelines emphasize 'transparency' and 'fairness,' highlighting the importance of recognizing and addressing societal biases that may be encoded in clinical models. The checklist integrates these concepts throughout, ensuring that bias and fairness are considered at every stage of the model's life cycle.
- **Modular Framework:** The new guidelines are modular, with different requirements based on the Research Design(s) and LLM Task(s) that are reported in a given study. This change was motivated by the wide variety of applications and approaches in biomedical LLM research, from model development through evaluation, necessitating more specialized reporting items.

# The TRIPOD-LLM Statement: A Targeted Guideline For Reporting Large Language Models Use

**Table 1.** Research design and LLM task categories for the modular TRIPOD-LLM guideline.

	Definition	Example
<b>Research design</b>		
<i>De novo</i> LLM development	Building a new language model from scratch or significantly fine-tuning existing base models to develop new functionalities or to adapt to new tasks.	A study pre-training a new LLM on a hospital's clinical data, e.g., Peng et al. (2023) <sup>51</sup>
LLM methods	Quantitative or theoretical investigations that focus on new architectures of model design, new computational methods to understand LLMs, new methods to evaluate LLMs, and/or new methods to optimize LLM prompting.	A study of a new retrieval-augmented generation LLM framework for medicine, e.g., Zakka et al. (2024) <sup>52</sup>
LLM evaluation	Assessing or testing an existing LLM to determine its efficacy, accuracy, or suitability for a specific task within healthcare may also include evaluating the risks and biases arising from using it.	A study investigating biased diagnostic reasoning in an existing LLM, e.g., Zack et al. (2023). <sup>21</sup>
LLM evaluation in healthcare settings	Evaluating an LLM when used as part of a clinical workflow, focusing on its integration and impact on clinical, administrative, or workforce outcomes.	A study reporting the performance of an LLM deployed in real-time to predict outcomes in hospitalized patients, e.g., Jiang et al. (2023) <sup>6</sup>
<b>LLM task</b>		
Text processing	Manipulating and lower-order processing of text data, which includes tasks including but not limited to tokenization, parsing, and entity recognition.	A study investigating a new LLM approach to named entity recognition, e.g., Keloth et al. 2024 <sup>53</sup>
Classification	Assigning predefined labels to text data.	A study fine-tuning an LLM to determine whether or not a sentence in a clinic note mentions one or several social determinants of health, e.g., Guevara et al. (2024) <sup>54</sup>
Long-form question-answering	Providing detailed answers to complex queries, which can involve reasoning over multiple documents or pieces of evidence. Note that multi-choice question-answering is subsumed under classification.	A study investigating the ability of an existing LLM to respond to patient portal messages, e.g., Chen et al., (2024) <sup>24</sup>
Information retrieval	The process of fetching relevant information from large datasets based on specific queries, which is relevant for tasks like literature review or patient history retrieval.	A study that trained a Transformer model to retrieve biomedical publications relevant to a user's query, e.g., Jin et al. (2023). <sup>55</sup>
Conversational agent (chatbot)	Responding and engaging in conversation with users, often used for patient interaction, health advisories, or as virtual assistants for healthcare providers.	A study investigating whether access to an interactive LLM-based chatbot impacts clinicians' diagnostic reasoning, e.g., Goh et al. (2024) <sup>56</sup>
Documentation generation	Automatically creating medical documentation from clinical data, dictations, or recordings.	A study evaluating the quality of clinical notes automatically generated from ambient clinic recordings, e.g., Tierney et al. (2024) <sup>5</sup>
Summarization and Simplification	Condensing large text documents into shorter versions or simplifying the content for easier comprehension is useful in patient education or in creating executive summaries of medical records.	A study evaluating the ability of LLMs to convert discharge summaries into patient-friendly plain language, e.g., Zaretsky et al. (2024) <sup>57</sup>
Machine translation	Converting text from one language to another.	A study comparing the ability of smaller language models fine-tuned for translation versus generalist LLMs to translate biomedical text between Spanish and English, e.g., Han et al. (2022) <sup>58</sup>
Outcome forecasting	Predicting future medical outcomes based on historical data, which can be used in prognosis estimation or treatment effectiveness studies.	A study investigating the ability of LLMs to predict out-of-hospital mortality in patients admitted to intensive care units, e.g., Yoon et al. (2024) <sup>59</sup>

## The TRIPOD-LLM Statement: A Targeted Guideline For Reporting Large Language Models Use

**Table 2:** TRIPOD-LLM Checklist

Section	Item	Checklist Item	Research Design	LLM Task
Title				
Title	1	Identify the study as developing, fine-tuning, and/or evaluating the performance of an LLM, specifying the task, the target population, and the outcome to be predicted.	All	All
Abstract				
Abstract	2	See TRIPOD-LLM for Abstracts	All	All
Introduction				
Background	3a	Explain the healthcare context / use case (e.g., administrative, diagnostic, therapeutic, clinical workflow) and rationale for developing or evaluating the LLM, including references to existing approaches and models.	All	All
	3b	Describe the target population and the intended use of the LLM in the context of the care pathway, including its intended users in current gold standard practices (e.g., healthcare professionals, patients, public, or administrators).	E H	All
Objectives	4	Specify the study objectives, including whether the study describes the initial development, fine-tuning, or validation of an LLM (or multiple stages).	All	All
Methods				
Data	5a	Describe the sources of data separately for the training, tuning, and/or evaluation datasets and the rationale for using these data (e.g., web corpora, clinical research/trial data, EHR data, or unknown).	All	All
	5b	Describe the relevant data points and provide a quantitative and qualitative description of their distribution and other relevant descriptors of the dataset (e.g., source, languages, countries of origin)	All	All
	5c	Specifically state the date of the oldest and newest item of text used in the development process (training, fine-tuning, reward modeling) and in the evaluation datasets.	All	All
	5d	Describe any data pre-processing and quality checking, including whether this was similar across text corpora, institutions, and relevant socio-demographic groups.	All	All
	5e	Describe how missing and imbalanced data were handled and provide reasons for omitting any data.	All	All
Analytical Methods	6a	Report the LLM name, version, and last date of training.	All	All
	6b	Report details of LLM development process, such as LLM architecture, training, fine-tuning procedures, and alignment strategy (e.g., reinforcement learning, direct preference optimization, etc.) and alignment goals (e.g., helpfulness, honesty, harmlessness, etc.).	M D	All

	6c	Report details of how text was generated using the LLM, including any prompt engineering (including consistency of outputs), and inference settings (e.g., seed, temperature, max token length, penalties), as relevant.	M D E	All
	6d	Specify the initial and post-processed output of the LLM (e.g., probabilities, classification, unstructured text).	All	All
	6e	Provide details and rationale for any classification and, if applicable, how the probabilities were determined and thresholds identified.	All	C OF
LLM Output	7a	Include metrics that capture the quality of generative outputs, such as consistency, relevance, and accuracy, compared to gold standards.	All	QA IR DG SS MT
	7b	Report the outcome metrics' relevance to downstream task at deployment time and, where applicable, correlation of metric to human evaluation of the text for the intended use.	E H	All
	7c	Clearly define the outcome, how the LLM predictions were calculated (e.g., formula, code, object, API), the date of inference for closed-source LLMs, and evaluation metrics.	E H	All
	7d	If outcome assessment requires subjective interpretation, describe the qualifications of the assessors, any instructions provided, relevant information on demographics of the assessors, and inter-assessor agreement.	All	All
	7e	Specify how performance was compared to other LLMs, humans, and other benchmarks or standards.	All	All
Annotation	8a	If annotation was done, report how text was labeled, including providing specific annotation guidelines with examples.	All	All
	8b	If annotation was done, report how many annotators labeled the dataset(s), including the proportion of data in each dataset that were annotated by more than 1 annotator, and the inter-annotator agreement.	All	All
	8c	If annotation was done, provide information on the background and experience of the annotators or characteristics of any models involved in labelling.	All	All
Prompting	9a	If research involved prompting LLMs, provide details on the processes used during prompt design, curation, and selection.	All	All
	9b	If research involved prompting LLMs, report what data were used to develop the prompts.	All	All
Summarization	10	Describe any preprocessing of the data before summarization.	All	SS
Instruction tuning/Alignment	11	If instruction tuning/alignment strategies were used, what were the instructions, data, and interface used for evaluation, and what were the characteristics of the populations doing evaluation?	M D	All
Compute	12	Report compute, or proxies thereof (e.g., time on what and how many machines, cost on what and how many machines, inference time, floating-point operations per second (FLOPs)), required to carry out	M D E	All

		methods.		
Ethical Approval	13	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent.	All	All
Open Science	14a	Give the source of funding and the role of the funders for the present study.	All	All
	14b	Declare any conflicts of interest and financial disclosures for all authors.	All	All
	14c	Indicate where the study protocol can be accessed or state that a protocol was not prepared.	H	All
	14d	Provide registration information for the study, including register name and registration number, or state that the study was not registered.	H	All
	14e	Provide details of the availability of the study data.	All	All
	14f	Provide details of the availability of the code to reproduce the study results.	All	All
Public Involvement	15	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement.	H	All
Results				
Participants	16a	When using patient/EHR data, describe the flow of text/EHR/patient data through the study, including the number of documents/questions/participants with and without the outcome/label and follow-up time as applicable.	E H	All
	16b	When using patient/EHR data, report the characteristics overall and, for each data source or setting, and for development/evaluation splits, including the key dates, key characteristics, and sample size.	E H	All
	16c	For LLM evaluation that include clinical outcomes, show a comparison of the distribution of important clinical variables that may be associated with the outcome between development and evaluation data, if available.	E H	All
	16d	When using patient/EHR data, specify the number of participants and outcome events in each analysis (e.g., for LLM development, hyperparameter tuning, LLM evaluation).	E H	All
Performance	17	Report LLM performance according to pre-specified metrics (see item 7a) and/or human evaluation (see item 7d).	All	All
LLM Updating	18	If applicable, report the results from any LLM updating, including the updated LLM and subsequent performance.	All	All
Discussion				
Interpretation	19a	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies.	All	All
Limitations	19b	Discuss any limitations of the study and their effects on any biases, statistical uncertainty, and generalizability.	All	All

Usability of the LLM in context	19c	Describe any known challenges in using data for the specified task and domain context with reference to representation, missingness, harmonization, and bias.	E H	All
	19d	Define the intended use for the implementation under evaluation, including the intended input, end-user, level of autonomy/human oversight.	E H	All
	19e	If applicable, describe how poor quality or unavailable input data should be assessed and handled when implementing the LLM, i.e., what is the usability of the LLM in the context of current clinical care.	E H	All
	19f	If applicable, specify whether users will be required to interact in the handling of the input data or use of the LLM, and what level of expertise is required of users.	E H	All
	19g	Discuss any next steps for future research, with a specific view to applicability and generalizability of the LLM.	All	All

LLM = large language model; M = LLM methods; D = *de novo* LLM development; E = LLM evaluation; H = LLM evaluation in healthcare settings; C = classification; OF = outcome forecasting; QA = long-form question-answering; IR = information retrieval; DG = document generation; SS = summarization and simplification; MT = machine translation; EHR = electronic health record.

Note: For studies using existing LLMs, users should include reference(s) to reportable information if provided by the original developers or state that this information is not available.



## The TRIPOD-LLM Statement: A Targeted Guideline For Reporting Large Language Models Use

**Table 3:** Essential items to include in a journal or conference abstract for a study describing the development, fine-tuning, or evaluation of a large language model (TRIPOD-LLM for Abstracts<sup>\*</sup>)

TRIPOD-LLM FOR ABSTRACTS				
Section	Item	Checklist Item	Research Design	LLM Task
Title	2a	Identify the study as developing, fine-tuning, and/or evaluating the performance of an LLM, specifying the task, the target population, and the outcome to be predicted.	All	All
Abstract	2b	Provide a brief explanation of the healthcare context, use case and rationale for developing or evaluating the performance of an LLM.	E H	All
Objectives	2c	Specify the study objectives, including whether the study describes LLMs development, tuning, and/or evaluation	All	All
Methods	2d	Describe the key elements of the study setting.	All	All
Methods	2e	Detail all data used in the study, specify data splits and any selective use of data.	M D E	All
Methods	2f	Specify the name and version of LLM(s) used.	All	All
Methods	2g	Briefly summarize the LLM-building steps, including any fine-tuning, reward modeling, reinforcement learning with human feedback (RLHF), etc.	M D	All
Methods	2h	Describe the specific tasks performed by the LLMs (e.g., medical QA, summarization, extraction), highlighting key inputs and outputs used in the final LLM.	All	All
Methods	2i	Specify the evaluation datasets/populations used, including the endpoint evaluated, and detail whether this information was held out during training/tuning where relevant, and what measure(s) were used to evaluate LLM performance.	All	All
Results	2j	Give an overall report and interpretation of the main results.	All	All
Discussion	2k	Explicitly state any broader implications or concerns that have arisen in light of these results.	All	All
Other	2l	Give the registration number and name of the registry or repository (if relevant).	H	All

LLM = large language model; M = LLM methods; D = *de novo* LLM development; E = LLM evaluation; H = LLM evaluation in healthcare settings