From text to treatment: the crucial role of validation for generative large language models in health care

Generative large language models (LLMs) have made incredible progress and are speculated to become the next big revolution in health care. Researchers have described several compelling uses for LLMs in health care, including the automatic generation of clinical information letters[1] and chatbots answering patient questions.[2-4] Thorough validation of developed LLMs is of the utmost importance to their safe and effective application in health-care practice, because incomplete LLM outputs or unchecked LLM hallucinations can be harmful to patient care. Yet, due to their generative nature there is no obvious translation of the LLM output into quantifiable outcomes.

The validation challenge is twofold. First, unlike the field of (quantitative) artificial intelligence (AI) prediction algorithms, whereby standardised validation criteria (such as discrimination using the c-index or area under the receiver operating characteristic curve) can be consistently applied across various tasks such as diagnosis and prognosis,[5] LLMs face a more complex landscape. Natural language generation (NLG) tasks vary substantially. For instance, validating a patient letter generated by a LLM requires a different validation approach compared with assessing responses from a chatbot. The former requires consistency with the electronic health record or other source document and the latter focusses on answer accuracy.

Second, in contrast to AI prediction models that typically produce a single numeric prediction (such as an outcome probability), LLMs exhibit remarkable versatility in their outputs. A single prompt might generate grammatically distinct yet equally valid text outputs. This extensive output space complicates the comparison of the output to a ground truth. Incorrectly generated output texts might be invalid in very distinct ways (eg, grammar, factuality, and completeness). This versatility in outputs makes it much harder to predict the clinical impact and unintended consequences of LLMs.

Two studies published in 2023 have highlighted concerning issues, such as LLMs perpetuating racial discrimination in health care[6] and providing harmful results when not moderated by health-care providers.[4] These examples underscore the critical need for continuous validation of LLM applications. Considering these complexities, we outline four key aspects for LLM validation which might be considered when assessing the validity of LLM outputs.

The first consideration is the choice of quality metrics and measurement. Due to the diversity in NLG tasks and LLM outputs, there is no uniform approach to defining validation metrics and measuring validation. This variability is apparent from the literature, where current LLM validation approaches can differ, even when applied to the same task. For example, ChatGPT's answers to medical questions have been evaluated on their accuracy with a binary metric (accurate or not),[2] but also evaluated on their empathy and quality on a five-point scale.[3] This variability of validation approaches could hinder the direct comparison of validation results of a particular LLM across studies and might lead to different conclusions about the validity of that LLM.

Moreover, some metrics do not have a clear definition—eg, the concept of humanness to rate the quality of patient letters on a scale from 0 to 10.[1] The choice and precise definition of quality metrics for LLM validation clearly warrant careful consideration. Researchers are advised to refer to existing guidance where applicable when deciding on the quality metrics and measurements for their LLM validation, for example human evaluation frameworks.[7]

The second consideration is the design of human evaluation studies. In the validation design of LLMs, a crucial decision lies between human and automatic validation. Although automatic validation metrics offer speed and efficiency, they are often poorly correlated with human evaluation scores.[8,9] Consequently, best practice dictates the use of human evaluations alongside automatic validation metrics. Several design questions should be addressed when validating LLMs based on human evaluations. Examples of design choices when evaluating and validating LLMs are the number of evaluators, the number of evaluations, and the averaging of different evaluator scores (eg, median, mean, or majority vote). Several frameworks for human evaluation methods exist that might provide guidance.[7,10]

Panel: **Three tiers of medical large language model (LLM) validation**

1. General validation
General validation assesses general LLM quality independent of the performed task. Important outcomes at this stage might be the LLM's robustness to different formulations of the same prompt and the readability of the LLM output.

2. Task specific validation
Task specific validation assesses the LLM performance on task specific outcomes. For example, for summarisation, the validation might focus on the consistency with source material and coverage of important clinical concepts.

3. Clinical impact validation
Clinical validation assesses the LLM performance and impact on specific health-care outcomes. The validation goals at this tier will depend on the clinical objectives and intended use, such as improved health outcomes, higher patient satisfaction, reduction in administration time, or improved workflows.

The third consideration is the validation of the intended use of the LLM. It is paramount to compare the generative outputs with current health-care practice to ensure the safe operation of LLMs for routine medical tasks. Current validation studies often employ a comparison of the LLM output versus human generated content. The intended use of many LLM applications envision a human verifying or adjusting the LLM output, which is often not incorporated into the validation design. It is advisable to align comparison studies with the intended use of the LLM, incorporating the human and LLM interaction where necessary. An example is provided by Chen and colleagues, who validated answers to patient questions produced by GPT-4 before and after physicians edited the draft responses.[4]

The fourth consideration is reporting of the validation results. Underreporting is pervasive throughout the biomedical literature and the LLM literature is not an exception. For instance, the exact prompt used to generate the LLM output is often not reported, even though changes in the prompt might lead to changes in output. Systematic reporting of all aspects of LLM validation will be greatly beneficial to the replicability and acceptability of these studies and researchers are recommended to adhere to existing reporting guidelines where possible, such as the SummEval for summary evaluation[9] or the announced chatbot assessment reporting tool.[11]

To move towards robust validation practices, we propose that the validation process for LLMs in health care should encompass three distinct tiers general, task specific, and clinical validation (panel). The evidence required at each tier might vary based on the intended use and the risks posed by the LLM in question.

The first tier of general validation focusses on assessing the overall robustness and quality of the LLM in the general target domain (eg, target language and clinical domain), regardless of the specific anticipated task. At this stage, aspects such as the LLM's responsiveness to slightly varied prompts and the fluency of the generated texts might be evaluated. Developers and researchers conducting feasibility checks for clinical applications will find this tier particularly relevant. A lack of general validity implies inadequate overall quality, rendering the LLM unreliable for health care contexts or human interpretation.

At the second tier, the task specific performance of the LLM is assessed. Validation practices will vary per task and focus on the quality of the generated text content. For instance, the consistency of the generated patient letters with the source material or the accuracy of the answers provide by a chatbot might be evaluated. Unintended (anticipated) consequences should be investigated, such as the risk of perpetuating racial biases through the LLM's output.[6] The task specific validation is important for both developers and end-users (patients and health-care providers). Task specific validation is directly related to the LLM's feasibility to perform the specific clinical task. A lack of task specific validity signals that the LLM is ill-suited for its intended purpose: it cannot be confidently relied upon to perform the required task satisfactorily.

In the third tier of validation, we examine the LLM's impact on health-care outcomes. The validation outcomes depend on the clinical objectives and the intended use of the LLM. These outcomes might include measurements related to patient outcomes improvement, reductions in administrative burden, or improved workflow of health-care professionals. This crucial stage directly involves the clinical end-users who will employ the LLM and provides them with a comprehensive understanding and evidence of how the tool can either enhance or pose risks to their existing health-care practices. Essentially, clinical validation answers the pivotal question: does the LLM effectively

contribute to desired changes in health-care practice? A lack of clinical validity signifies that despite technical soundness, the LLM falls short of achieving the desired clinical impact.

The launch of large language models such as ChatGPT has produced a spike in experimentation and use of this technology in health-care settings. Recent studies have shown great promise for the application of LLMs in health care[1–3] but often lacked the methodological rigor to ensure their safe and effective use in health-care practice. The variation in validation approaches currently observed for LLMs is understandable, considering the novelty of LLM techniques and absence of available guidance. Although variation in validations is not inherently problematic, it becomes an issue when a validation study is suboptimally designed, impeding replicability and hampering widespread application of LLMs in health-care practice. Robust validation practices can help, assessing general, task specific, and clinical validity. To ensure their safe performance over time, LLMs will have to be repeatedly validated even after implementation. LLMs might only live up to their promise for the health-care sector through replicable and repeated validation practices and regular updating of these practices in this new and upcoming field of interest.

We declare no competing interests. TL declares funding from the Dutch Research Council, outside this publication.

*Anne de Hond, Tuur Leeuwenberg, Richard Bartels, Marieke van Buchem, Ilse Kant, Karel GM Moons, Maarten van Smeden
a.a.h.dehond@umcutrecht.nl

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, 3584 CG Utrecht, Netherlands (AdH, TL, RB, KGMM, MvS); Clinical AI Implementation and Research Lab, Leiden University Medical Center, Leiden, Netherlands (MvB); Department of Information Technology and Digital Innovation, Leiden University Medical Center, Leiden, Netherlands (MvB); Department of Digital Health, University Medical Centre Utrecht, Utrecht, Netherlands (RB, IK)

1 Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. *Lancet Digit Health* 2023; **5:** e179–81.
2 Johnson SB, King AJ, Warner EL, Aneja S, Kann BH, Bylund CL. Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information. *JNCI Cancer Spectr* 2023; **7:** pkad015.
3 Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023; **183:** 589–96.
4 Chen S, Guevara M, Moningi S, et al. The effect of using a large language model to respond to patient messages. *Lancet Digit Health* 2024; published online April 24. https://doi.org/10.1016/S2589-7500(24)00060-8.
5 van Smeden M, Moons KGM, Hooft L, Chavannes N, van Os HJA, Kant I. Guideline for high-quality diagnostic and prognostic applications of AI in healthcare. 2023. https://guideline-ai-healthcare.com (accessed April 15, 2024).
6 Omiye JA, Lester JC, Spichak S, Rotemberg V, Daneshjou R. Large language models propagate race-based medicine. *NPJ Digit Med* 2023; **6:** 195.
7 van der Lee C, Gatt A, Van Miltenburg E, Wubben S, Krahmer E. Best practices for the human evaluation of automatically generated text. 12th International Conference on Natural Language Generation. Oct 29–Nov 1, 2019 (W19-8643).
8 Gehrmann S, Clark E, Sellam T. Repairing the cracked foundation: a survey of obstacles in evaluation practices for generated text. *J Artif Intell Res* 2023; **77:** 103–66.
9 Fabbri AR, Kryściński W, McCann B, Xiong C, Socher R, Radev D. Summeval: Re-evaluating summarization evaluation. *Trans Assoc Comput Linguist* 2021; **9:** 391–409.
10 van der Lee C, Gatt A, van Miltenburg E, Krahmer E. Human evaluation of automatically generated text: current trends and best practice guidelines. *Comput Speech Lang* 2021; **67:** 101151.
11 Huo B, Cacciamani GE, Collins GS, McKechnie T, Lee Y, Guyatt G. Reporting standards for the use of large language model-linked chatbots for health advice. *Nat Med* 2023; **29:** 2988.