BEYOND THE FACTUAL VS. HALLUCINATORY DICHOTOMY: A REFINED TAXONOMY FOR LLM MEDICAL RESPONSE CATEGORIZATION

Saleh Afroogh†*, Yasser Pouresmaeil†, Junfeng Jiao
Urban Information Lab, School of Architecture
The University of Texas at Austin, Austin, TX, USA
IPM Institute for Research in Fundamental Sciences (PhD alumnus), Tehran, Iran
saleh.afroogh@utexas.edu

ABSTRACT

Large Language Models (LLMs) are increasingly used in medicine, but the traditional factual/hallucinatory distinction fails to reflect the evolving nature of medical knowledge. This paper critiques that binary and proposes a refined, three-tiered classification: (1) Currently Verifiable Responses, (2) Tentatively Examinable Responses, and (3) Predictive Responses. This framework introduces a veridicality gradient and emphasizes temporal verifiability, enabling more accurate evaluation, reducing clinical risk, and supporting adaptive model calibration. Ultimately, it promotes the development of safer and more epistemically responsible medical AI systems.

Keywords: LLMs in medicine, AI veridicality, established medical knowledge, predictive responses, provisional knowledge, AI safety in healthcare, temporal verifiability.

1 INTRODUCTION: PROMISE AND PERIL OF LLMS IN MEDICINE

Large Language Models (LLMs) are increasingly used in medicine, healthcare, and medical research, including clinical support and patient education to evidence syntheses generation and assistance in biomedical research (Ahmad et al., 2023; Qiu et al., 2024; Lee & Lindsey, 2024; Nassiri & Akhloufi, 2024; Williams et al., 2024; He et al., 2024). However, the growing reliance on LLMs in sensitive domains has drawn attention to the necessity of a thorough testing of the veridicality of their responses—i.e., the extent to which their response aligns with up-to-date medical knowledge and scientific fact. Traditionally, LLM responses have been put into a binary category: first would be the *Factual Responses* that refer to outputs that align with established, proven medical knowledge. Secondly, in contrast, *Hallucinatory Responses* are fabricated, false, or misleading outputs that lack a factual source, though they may be presented as accurate (Pal et al., 2023; Kim et al., 2025; Agarwal et al., 2024; Gu et al., 2024).

While this dichotomy is helpful as a primitive safeguard against disinformation (Pham & Vo, 2024; Asgari et al., 2024; Hegselmann et al., 2024b; Priola, 2024; Chen et al., 2024; Hatem et al., 2023), and while varieties of hallucination in medical LLMs have been distinguished (Vishwanath et al., 2024), it oversimplifies the dynamic, probabilistic, and context-dependent character of medical knowledge, as such knowledge is not static but is in a state of continuous evolution over time by way of research, clinical trials, and paradigm shifts. Consequently, the majority of LLM responses fall into a gray zone of tentative or predictive knowledge that the binary distinction cannot capture.

*† Equal contribution. * Corresponding author.

2 THE LIMITATIONS OF THE FACTUAL/HALLUCINATORY DICHOTOMY

2.1 PRESUMPTION OF STATIC VERIDICALITY

The fact/hallucination framework tacitly assumes that any medical answers can be granted a veridicality status of some kind when generated. This assumption, however, ignores the evolving or developing nature of scientific facts. Medical knowledge usually advances in waves of initial acceptance, zealous confirmation, and eventual agreement—or sometimes rejection. For instance, early guesswork about novel biomarkers could receive early acceptance but is amenable to eventual confirmation (Califf, 2018), or new drug treatments may appear effective at first but require long-term evidence of safety prior to being substantiated in clinical practice guidelines (Nature Medicine, 2025). In placing LLM responses these either on the category of fact or hallucination, we risk reductionism on a continuum along which most responses are epistemically indeterminate or non-examinable, and not true or false.

2.2 FAILURE TO ACCOUNT FOR PREDICTIVE STATEMENTS

Another shortcoming of the classical dichotomy is its inability to accommodate predictive responses. Predictive assertions—such as predictions of disease course, survival probability, or treatment outcome—can rely on probabilistic models, clinical trends, and evidenced risk factors. Whereas some predictive responses are amenable to test against conventional probabilistic models or standards of medical methodology, there are unexplainable LLM predictions that have to await subsequent empirical findings to determine their validity. Treating such responses as either true or hallucinatory fails to take into account their prospective status, in which accuracy can be determined only in hindsight.

2.3 BLIND SPOT TO PROVISIONAL KNOWLEDGE

The fact/hallucination model also ignores the provisional nature of early scientific knowledge. Some solutions, though not yet established, are consonant with the best current evidence or dominant clinical paradigms. Others may have speculative or incorrect statements that are not consonant with dominant state of provisional knowledge. Without a mechanism to recognize these cases, we risk conflating scientific skepticism with epistemic untrustworthiness.

3 SUGGESTED REFINEMENT: THREE-LEVEL CLASSIFICATION OF LLM MEDICAL RESPONSES

In order to overcome these limitations, we propose a three-level classification scheme that offers greater epistemic granularity:

3.1 CURRENTLY VERIFIABLE RESPONSES (SETTLED KNOWLEDGE)

By verifiable responses we mean those to which veridicality can be established with high confidence by appeal to available resources, tools, and knowledge. In this framework, factual responses are those in agreement with established, known medical facts, while hallucinatory responses are those that contradict or deviate from available medical facts, such as recommended dosage levels for frequently prescribed drugs and established diagnostic characteristics for frequent diseases.

3.2 PROVISIONALLY EXAMINABLE RESPONSES (PROVISIONAL/CONTINGENT KNOWLEDGE)

There are LLM responses that are provisionally examinable. These are the ones citing knowledge or hypotheses that are not yet testable by means of currently available resources but potentially confirmed or falsified in the future. There are two subcategories of such responses: (1) *Provisional-Accurate*: those adhering to best existing evidence or current scientific consensus, and (2) *Provisional-Inaccurate*: those contradictory to provisional knowledge at the moment, and potentially resulting in epistemic flaws. Examples include early findings of a relationship between gut microbiota and mental health, awaiting larger-scale corroboration (Xiong et al., 2023), and speculations about the long-term impact of mRNA vaccines on immune modulation (Khoury et al., 2021).

3.3 Predictive Responses (Prospective/Probabilistic Statements)

Finally, predictive responses by LLMs are utterances yielding forward-looking probabilistic predictions about future events or outcomes. There are two veridicality aspects to predictive LLM responses: (1) in terms of their content—the proposition that is predicted, and (2) in terms of the soundness of the prediction method or procedure.

As for (1), we need to wait for the outcome in order to see whether the predictive statement was indeed true or false. However, regarding (2), sometimes such predictive responses are explainable in terms of predictive methodologies in medical research, which might be valid if it aligns with the conventional standards used for such predictions, or invalid if it does not. At other times, such responses are not explainable at all, resulting from the peculiar ways in which LLMs may identify patterns in their training data, in which case there is no way to assess the validity of such prediction.

Note that the validity of a prediction differs from whether or not its content proves true or false. Examples include estimates of the remission rate for a therapy for cancer based on acknowledged clinical data, and forecasts about the eventual success of a new therapy under ongoing clinical trials.

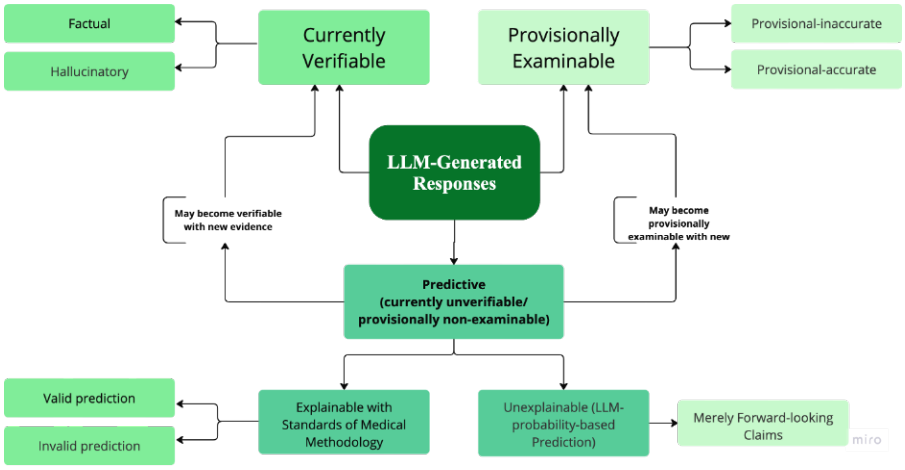

Figure 1: Classification of LLM-Generated Medical Responses

4 Refining the Classification of Provisionally Examinable Responses: An Epistemic Gradient

One particularly crucial refinement involves the inclusion of a veridicality gradient in the provisionally examinable category with *Provisional-Accurate* and *Provisional-Inaccurate* responses being distinguished. Provisional-Accurate Responses align with the provisional state of affairs of knowledge and are the hypotheses or preliminary outcomes most likely, while Provisional-Inaccurate Answers diverge from or misrepresent prevailing paradigms, possibly introducing epistemic mistake or perplexity. This distinction allows reviewers to assess coherence with new evidence even in the case of a response's terminal veridicality being open to doubt.

4.1 The Role of Temporal Verifiability and Epistemic Certainty

To operationalize this subtle scheme, answers can be evaluated along two axes: *temporal verifiability*, which concerns whether a response can be verified at the time of generation or requires future empirical evidence; and *epistemic certainty*, which addresses whether the response is grounded in an established body of knowledge or situated within an evolving and tentative paradigm.

One of the main disadvantages of current AI evaluation frameworks is that they fail to take into consideration the temporal dynamics of medical knowledge. Medical science is dynamic, and established clinical guidelines and treatment methods can be revised or refuted in the future (Williams et al., 2024; Vishwanath et al., 2024).

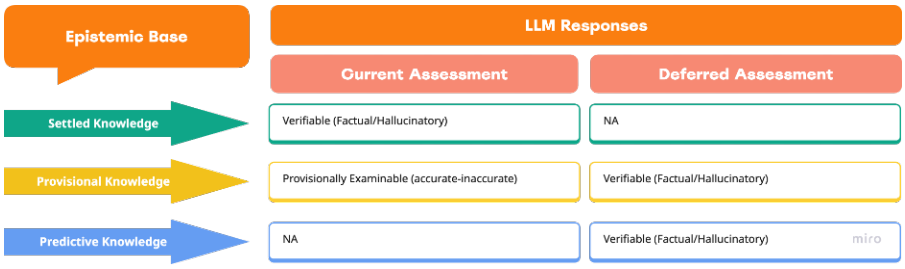

Figure 2: Temporal Verifiability and Epistemic Uncertainty

A time-conscious AI assessment system is required to:

- Recertify AI-made medical claims when fresh scientific evidence becomes available.
- Make a distinction between incorrect but previously correct answers and actual hallucinations.
- Leverage real-time knowledge retrieval systems to allow LLMs to provide up-to-date medical information (Hegselmann et al., 2024a).

To enable AI trustworthiness in clinical settings, LLMs must be designed to refresh their knowledge base dynamically and provide an explicit declaration regarding the temporal validity of their outputs. Confidence scores and uncertainty labels must be ingrained in AI-generated predictions so that clinicians can appropriately interpret probabilistic predictions rather than embracing them as absolute clinical recommendations (Freyer et al., 2024).

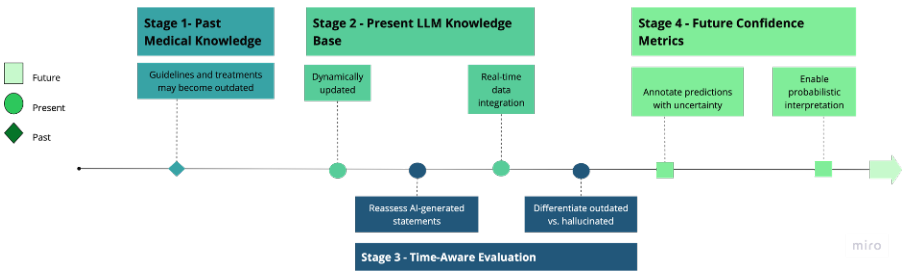

Figure 3: Time-Sensitive Evaluation Framework for LLMs in Healthcare

4.2 Implications for Medical AI Evaluation and Oversight

Improved Accuracy Assessment: The enriched classification provided by provisional and predictive responses makes it possible to more finely measure accuracy beyond binary correctness. By assessing alignment with provisional knowledge as well as predictive verifiability, medical AI systems can be tested more rigorously.

Risk Reduction in High-Stakes Applications: Provisional and predictive LLM responses as distinct categories prevent the miscription of rough-and-ready knowledge as definite fact—a vital protection against poor clinical decision-making.

Facilitating Adaptive Learning and Calibration: An augmented framework provides regular calibration of LLMs by observing the manner in which preliminary knowledge develops into certain knowledge with time, thereby allowing the model responses to adapt to remain in conformity with evolving paradigms in medicine.

5 Conclusion: Towards a More Sophisticated and Responsible AI in Medicine

As LLMs play a growing role in shaping the future of medicine, epistemic sensitivity in classifying their responses becomes crucial. Overcoming the simplistic factual/hallucinatory dichotomy is required to grasp the richness of medical knowledge, including anticipatory and contingent responses, and ultimately to enable the safety, reliability, and clinical utility of AI-generated medical insights. Through an augmented paradigm of classification, we may look to the future in which LLMs in medicine work not as information vaults but as active, contextual guides capable of mapping the moving landscape of scientific understanding with precision.

Conflict of Interest

The authors declare that the investigationation was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

In accordance with MLA (Modern Language Association) guidelines, we note the use of AI-powered tools, such as OpenAI's applications, for assistance in editing and brainstorming.

References

V. Agarwal et al. Medhalu: Hallucinations in responses to healthcare queries by large language models. *arXiv preprint arXiv:2409.19492*, 2024. doi: 10.48550/arXiv.2409.19492.

M. A. Ahmad, I. Yaramis, and T. D. Roy. Creating trustworthy llms: Dealing with hallucinations in healthcare ai. *arXiv preprint arXiv:2311.01463*, 2023. doi: 10.48550/arXiv.2311.01463.

E. Asgari et al. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. 2024. doi: 10.1101/2024.09.12.24313556.

R. M. Califf. Biomarker definitions and their applications. *Experimental Biology and Medicine*, 243(3):213–221, 2018. doi: 10.1177/1535370217750088.

J. Chen et al. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*, 2024. doi: 10.48550/arXiv.2406.10185.

O. Freyer, I. C. Wiest, J. N. Kather, and S. Gilbert. A future role for health applications of large language models depends on regulators enforcing safety standards. *Lancet Digital Health*, 6(9):e662–e672, 2024. doi: 10.1016/S2589-7500(24)00124-9.

Z. Gu, C. Yin, F. Liu, and P. Zhang. Medvh: Towards systematic evaluation of hallucination for large vision language models in the medical context. *arXiv preprint arXiv:2407.02730*, 2024. doi: 10.48550/arXiv.2407.02730.

R. Hatem, B. Simmons, and J. E. Thornton. A call to address ai 'hallucinations' and how healthcare professionals can mitigate their risks. *Cureus*, 2023. doi: 10.7759/cureus.44720.

K. He et al. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*, 2024. doi: 10.48550/arXiv.2310.05694.

S. Hegselmann, S. Z. Shen, F. Gierse, M. Agrawal, D. Sontag, and X. Jiang. A data-centric approach to generate faithful and high quality patient summaries with large language models. *arXiv preprint arXiv:2402.15422*, 2024a. doi: 10.48550/arXiv.2402.15422.

S. Hegselmann et al. Medical expert annotations of unsupported facts in doctor-written and llm-generated patient summaries. *PhysioNet*, 2024b.

J. Khoury et al. Covid-19 vaccine – long term immune decline and breakthrough infections. *Vaccine*, 39(48): 6984–6989, 2021. doi: 10.1016/j.vaccine.2021.10.038.

Y. Kim et al. Medical hallucination in foundation models and their impact on healthcare. 2025. doi: 10.1101/2025.02.28.25323115.

S. A. Lee and T. Lindsey. Can large language models abstract medical coded language? *arXiv preprint arXiv:2403.10822*, 2024. doi: 10.48550/arXiv.2403.10822.

K. Nassiri and M. A. Akhloufi. Recent advances in large language models for healthcare. *BioMedInformatics*, 4(2):1097–1143, 2024. doi: 10.3390/biomedinformatics4020062.

Nature Medicine. The future of medicine. *Nature Medicine*, 31(1):1–1, 2025. doi: 10.1038/s41591-024-03464-y.

A. Pal, L. K. Umapathi, and M. Sankarasubbu. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*, 2023. doi: 10.48550/arXiv.2307.15343.

D. K. Pham and B. Q. Vo. Towards reliable medical question answering: Techniques and challenges in mitigating hallucinations in language models. *arXiv preprint arXiv:2408.13808*, 2024. doi: 10.48550/arXiv.2408.13808.

M. P. Priola. Addressing hallucinations with rag and nmiss in italian healthcare llm chatbots. *arXiv preprint arXiv:2412.04235*, 2024. doi: 10.48550/arXiv.2412.04235.

J. Qiu, W. Yuan, and K. Lam. The application of multimodal large language models in medicine. *Lancet Regional Health - Western Pacific*, 45:101048, 2024. doi: 10.1016/j.lanwpc.2024.101048.

P. R. Vishwanath et al. Faithfulness hallucination detection in healthcare ai. In *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*. 2024. Available: `https://openreview.net/forum?id=6eMIzKFOpJ`.

C. Y. K. Williams et al. Evaluating large language models for drafting emergency department discharge summaries. 2024. doi: 10.1101/2024.04.03.24305088.

R.-G. Xiong et al. The role of gut microbiota in anxiety, depression, and other mental disorders as well as the protective effects of dietary components. *Nutrients*, 15(14):3258, 2023. doi: 10.3390/nu15143258.