Journal of Digital Health
https://ojs.luminescience.cn/JDH

LUMINESCIENCE
https://luminescience.cn

Original research

Prompt style effects on large language model differential diagnosis output in adolescent catatonia and psychosis

Thomas F. Heston[1,2]

[1] Department of Family Medicine, University of Washington, Seattle, Washington USA
[2] Department of Medical Education and Clinical Sciences, Washington State University, Spokane, Washington USA

Corresponding to: Thomas F. Heston, Email: theston@uw.edu

Abstract: Prompt engineering shapes large language model output, but practicing clinicians have little guidance on which prompting style best suits a given clinical task. In an exploratory single-case design, one publicly available, de-identified case report of a 15-year-old girl who presented with catatonia and psychosis following high-dose ashwagandha ingestion was submitted to a base, unauthenticated ChatGPT-5.5 session under four prompt conditions: zero-shot, chain-of-thought, role/persona, and structured/constrained output. Outputs were scored from 0 (feature absent) to 2 (complete) on four criteria: diagnostic accuracy, causative-agent recognition, dangerous-mimic recognition, and calibration. Output was compared with the published clinical course. Total scores were zero-shot 7/8, chain-of-thought 6/8, role/persona 7/8, and structured 7/8. All four conditions ranked bipolar I disorder above substance-induced psychotic disorder, whereas the treating team's documented differential favored unspecified and substance-induced categories without a confirmed bipolar diagnosis. All four conditions recommended neuroimaging, electroencephalography, and lumbar puncture, exceeding the workup performed by the treating team. In this case, core diagnostic content was materially the same across the four prompt styles, which differed in probability quantification and reasoning transparency rather than in the underlying differential diagnosis generated. These observations are hypothesis-generating and specific to this vignette and model version; replication across cases and models is required before generalization.

Keywords: Prompt engineering, Large language model, Differential diagnosis, Clinical decision support, Catatonia, Ashwagandha

1. Introduction

Large language models (LLMs) are increasingly consulted by clinicians for diagnostic reasoning support, yet the wording and structure of the prompt, not only the underlying model, materially shape the output returned [1-4]. Prior work has described prompt-engineering principles for medical education and healthcare applications [2, 5-7] and has evaluated selected prompting strategies in clinical natural language processing and diagnostic-vignette settings [3, 8, 9]. Chain-of-thought reasoning has been shown to mitigate some cognitive biases across several LLMs in vignette-based tasks, suggesting that reasoning structure can affect substantive output characteristics rather than merely the response format [10-11].

Studies have examined prompt effects specifically within psychiatry. In one, the open-ended question “How

Received: Jul. 10, 2026; Revised: Jul. 29, 2026; Accepted: Sep. 3, 2026; Published: Nov. 2, 2026
Copyright © 2026 Thomas F. Heston.
DOI: https://doi.org/10.55976/jdh.5202616146-12
This is an open-access article distributed under a CC BY license (Creative Commons Attribution 4.0 International License)
https://creativecommons.org/licenses/by/4.0/

can I feel happier?" was asked in the context of a stated history of depression, and submitted to ChatGPT-4.0 under four prompt variations distinguished mainly by assigned role; responses differed substantially, and one persona produced suggestions characterized as potentially harmful [12]. In another, both model choice and prompt specificity affected the quality of psychiatric multiple-choice questions generated for assessment purposes [13]. Both examined generative tasks, in which the model composes advice or new test items and the space of acceptable outputs is wide. Neither examined a constrained diagnostic task, in which the model must rank a differential diagnosis against an actual clinical presentation and recommend a workup. Another recent study addressed constrained psychiatric diagnosis: integrating expert-derived decision-tree reasoning into the prompt substantially improved positive predictive value and reduced overdiagnosis on Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR) vignettes while preserving most true-positive detections [14]. Whether simpler variations in prompt style alone exert the same influence on constrained diagnostic ranking that they exert on open-ended generation therefore remains only partially established.

The distinction has clinical and educational consequences. If prompt style materially changes the differential a model produces, then clinicians consulting an LLM at the point of care require training in prompt construction before the output can be relied upon, and medical curricula must treat prompting as a clinical skill in its own right [7, 15]. If prompt style instead changes only how the same differential is presented, the training priority shifts from prompt construction toward critical appraisal of model output. These two conclusions imply different educational investments, and the question is currently unresolved for diagnostic tasks.

This study tests four common prompt styles—zero-shot [8], chain-of-thought [9], role/persona [12], and structured/constrained output [3]—on a single published pediatric psychiatric case to determine whether the diagnostic content produced differs by prompt style, and, if so, in what respects.

2. Methods

2.1 Case source

The case vignette was adapted from a published, open-access case report of a 15-year-old female who presented with catatonia and psychosis following high-dose ashwagandha ingestion, published under a Creative Commons Attribution (CC BY) license [16]. That license permits adaptation and redistribution with attribution, so no separate permission from the authors or publisher was required, and attribution is provided by citation. The vignette was de-identified and included all history, examination findings, and laboratory results available to the treating team at the time of psychiatric consultation. Withheld from the submitted vignette were the article title, the published discussion, the stated conclusion, and the discharge diagnosis, each of which names or implies a diagnostic formulation, so that model output could be compared with the treating team's reasoning rather than restate it.

2.2 Model

A single LLM, ChatGPT-5.5 (Open artificial intelligence (Open AI), San Francisco, CA, USA), was queried in July 2026 through the public consumer web interface. Each prompt condition was submitted as the opening message of a separate, newly opened session, and every session was unauthenticated and conducted in an incognito browser window to avoid the influence of stored account history, memory features, or preceding conditions on output. ChatGPT-5.5 was selected because it was the default model served to general users at the time of the study, and therefore the version a clinician querying ChatGPT would most likely encounter. The consumer interface does not expose decoding parameters; temperature, top-p, sampling seed, and any vendor system prompt are provider-set, not user-configurable, and not publicly disclosed. Output is therefore non-deterministic, and each condition was run once. Reporting follows the Transparent Reporting of a Multivariable Model for Individual Prognosis or Diagnosis (TRIPOD)-LLM guideline for studies using large language models [17].

2.3 Training-data contamination

The source case report is openly accessible, so the possibility that the model encountered it during training requires consideration. The vendor documents a training-data cutoff of December 1, 2025, for this model version [18], and the source article was published in 2026, so the case report could not have been in the training corpus. Therefore, training-data contamination did not occur.

2.4 Prompt conditions

The identical case vignette and closing question ("What is your differential diagnosis, ranked by likelihood? What workup would you recommend?") were submitted under four conditions that differed only in the instructional framing appended to the prompt. The appended text is reproduced verbatim below, and the complete submitted prompts and unedited model outputs are available in the repository identified in the Data Availability Statement.

1. Zero-shot: no instruction beyond the closing question, restated once as "What is your differential diagnosis, ranked by likelihood, with recommended workup?"

2. Chain-of-thought: "List the key clinical features first,

then reason through each diagnostic possibility before giving your ranked differential."

3. Role/persona: "You are a board-certified child and adolescent psychiatrist evaluating this patient in the ED." (ED, emergency department).

4. Structured/constrained output: "For each diagnosis, provide: likelihood (%), supporting features, refuting features, and next diagnostic step."

2.5 Evaluation rubric

The four criteria were selected to capture distinct failure modes of LLM diagnostic output: arriving at the wrong leading diagnosis, overlooking the precipitating agent, missing a medical condition capable of mimicking the presentation, and expressing unwarranted certainty. The rubric was developed specifically for this study, rather than adapted from a validated instrument, and it has not undergone psychometric evaluation. It therefore functions as a structured scoring aid rather than a measurement scale, and the resulting totals should be read as descriptive summaries rather than quantitative estimates of performance. Each output was scored by the author on the four criteria, each rated 0 (absent), 1 (partial), or 2 (complete), for a maximum score of 8:

• Diagnostic accuracy: whether the differential included the diagnosis reflected in the treating team's documented course (substance-induced psychotic disorder, not a confirmed bipolar diagnosis), and where it ranked.

• Causative-agent recognition: whether ashwagandha was identified as a plausible precipitant with a stated mechanism [19], rather than mentioned only in the background history.

• Dangerous-mimic recognition: whether the output identified the medical conditions capable of mimicking a primary psychiatric presentation and requiring exclusion, principally anti-N-methyl-D-aspartate receptor encephalitis [20], and specified a diagnostic plan able to exclude them. This criterion was deliberately not scored by whether the recommended workup exceeded the studies the treating team performed. Recommending more testing is not equivalent to recommending better testing, and a broader investigation may represent low-value or defensive care rather than superior safety [21].

• Calibration: whether the output expressed appropriate diagnostic uncertainty (a ranked or probabilistic differential) rather than a single, falsely confident conclusion.

• Scoring was performed by the author, a physician board-certified in family medicine and nuclear medicine, against the published case course. A single rater was used by design: the study measures the effect of prompt style, and additional raters would introduce interrater variance into a comparison in which every other element was held constant. Blinding was not possible, since the output format itself identifies the condition. The criteria were fixed before any output was read, and all four outputs are deposited in full, so any reader can verify the scoring against the source text.

3. Results

All four prompt conditions produced a complete ranked differential and a recommended workup. Rubric totals were 7 of 8 for zero-shot, role/persona, and structured/constrained output, and 6 of 8 for chain-of-thought (Table 1). All four scored identically on diagnostic accuracy, causative-agent recognition, and dangerous-mimic recognition, differing only in calibration.

Table 1. Rubric scores by prompt condition (0-2 per criterion, maximum 8)

Criterion	Zero-shot	Chain-of-thought	Role/persona	Structured
Diagnostic accuracy	1	1	1	1
Causative-agent recognition	2	2	2	2
Dangerous-mimic recognition	2	2	2	2
Calibration	2	1	2	2
Total (/8)	7	6	7	7

All four conditions produced a ranked differential headed by bipolar I disorder, manic episode with psychotic features and catatonia, citing the strong maternal history of bipolar I disorder with psychotic and catatonic features as the primary supporting factor. Substance/supplement-induced bipolar or psychotic disorder was uniformly ranked second and described as a plausible precipitant of a first episode in a genetically susceptible patient rather than the primary etiology. This ranking diverges from the treating team's documented differential, which listed unspecified psychotic disorder, substance (ashwagandha)-induced psychotic disorder, generalized anxiety disorder,

and unspecified depressive disorder, without a confirmed bipolar diagnosis [16]. That documented differential serves throughout as a clinical comparator rather than as an adjudicated ground truth. It records the formulation reached by treating clinicians at one point in the patient's course, not a verified final diagnosis established by independent expert review, so divergence from it reflects disagreement rather than demonstrated model error.

All four conditions recommended brain magnetic resonance imaging, electroencephalography, and, depending on the clinical course, lumbar puncture with an autoimmune encephalitis antibody panel to exclude anti-N-methyl-D-aspartate receptor encephalitis. The treating team's actual workup was limited to urine toxicology, a heavy metal panel, thyroid studies, a basic metabolic panel, and hematologic studies [16]; none of the four prompt-generated outputs matched this narrower scope, and all exceeded it.

The chain-of-thought output was the only condition that did not assign numeric likelihood percentages to individual diagnoses, using qualitative terms such as "very high" and "moderately likely" instead. It produced the longest and most granular feature-by-feature reasoning trail of the four conditions, listing eleven distinct diagnostic possibilities compared with eight to ten in each of the other three conditions. The structured condition returned explicit percentage estimates and used a consistent supporting-features/refuting-features format for each diagnosis. The role/persona condition returned percentage estimates for the leading diagnoses and included supporting and opposing considerations, but used a less uniform format across the full differential. All four conditions expressed graded uncertainty rather than a single confident conclusion. The three conditions assigning numeric likelihoods scored 2; the chain-of-thought condition scored 1, since ordinal labels such as "very high" convey rank but not magnitude and are interpreted inconsistently by different readers [22]. What differed was how that uncertainty was expressed, from ordinal verbal labels in the chain-of-thought condition to point estimates summing to 100% in the structured condition. This is a difference in the form of the expression, rather than in whether uncertainty was conveyed.

4. Analysis

The primary finding is that, in this case, prompt style altered the precision, structure, and presentation of the output more than it altered the leading diagnostic content. All four conditions converged on bipolar I disorder with psychotic and catatonic features as the leading diagnosis, an ashwagandha-associated substance- or supplement-induced mood or psychotic disorder as the principal alternative, and autoimmune encephalitis as the key dangerous mimic to exclude. Where the conditions diverged was in reasoning transparency, probability quantification, clinical register, and the breadth of lower-ranked diagnoses, rather than in the leading diagnostic formulation.

5. Discussion

This has a direct clinical implication: for the practicing clinician, these results suggest that prompt selection functions less as a lever on diagnostic content and more as a lever on how that content is packaged for a specific use. Structured or constrained prompting, which returns an itemized list of likelihoods, supporting features, refuting features, and next steps, may be well suited to documentation, such as drafting a chart-note differential that a clinician can then edit [23-24]. Chain-of-thought prompting, which provides a visible, stated rationale rather than a scored table, may be most useful for teaching or for clinicians who want to inspect the model's articulated reasoning before deciding how much weight to give its conclusion, consistent with prior evidence that chain-of-thought prompting can alter substantive clinical-vignette outputs, including bias patterns [10, 25]. Role or persona prompting did not change the diagnostic content relative to zero-shot in this case, but it returned a more clinically idiomatic tone; persona components can shift proposed testing and diagnostic breadth [26]. Zero-shot prompting, requiring the least construction effort, produced output that was materially equivalent to that of the more elaborate conditions in this case, suggesting it is a reasonable default for well-specified clinical questions with a capable model [27].

The consistent ranking of bipolar I disorder above substance-induced etiology across all four conditions is notable, independent of prompt style. All four outputs cited the maternal history as the dominant supporting factor, whereas the treating clinicians favored unspecified and substance-induced categories pending the longitudinal course. One possible explanation is that the model gave greater weight to a striking first-degree family history than the treating team did, rather than integrating it as one factor among several [28]. This hypothesis arises from a single case, not a demonstrated failure mode, and testing it would require replication across cases in which family history and presenting features point in different directions [29]. Pending such replication, a clinician using any of these four prompt styles would reasonably treat family-history-driven ranking as a hypothesis to weigh against the full clinical picture rather than as a conclusion.

This is consistent with prior work showing that LLMs can generate clinically sound content while remaining unreliable at converging on a single correct answer. ChatGPT-4 has been shown to produce cardiac risk stratifications that correlate closely with validated scores such as the Thrombolysis in Myocardial Infarction (TIMI) and History, Electrocardiogram, Age, Risk factors,

and Troponin (HEART), yet it gives inconsistent risk assessments when presented with identical chest pain vignettes on separate occasions [30]. A United Kingdom (UK) replication of a human–AI collaboration trial found that physicians using an LLM scored 21 percentage points below the model working alone, an effect attributed largely to clinicians querying the model in only 30% of applicable case questions rather than to weaker standalone model reasoning [31]. A recent perspective argues that text-based artificial intelligence can now reason at a level comparable to a physician, and that the primary challenge has shifted from raw capability to safe clinical implementation, including governance and workflow accountability [32]. Taken together with the present finding, in which all four prompt styles converged on the same well-reasoned differential but also assigned the highest rank to bipolar I disorder relative to the treating team's documented differential, this literature supports treating current-generation LLMs as strong generators of a differential diagnosis while withholding reliance on them to determine the single final diagnosis.

6. Limitations

This is an exploratory single-case study: one vignette, one model, one time point. The findings may not generalize to other diagnoses or model families; holding all three constant is nonetheless what isolates prompt style as the only variable. Scoring was performed by one rater, which holds rater variance constant for the same reason, and the criteria were applied to deposited outputs that any reader can verify independently. No statistical comparisons were attempted, which is appropriate given four observations. Few-shot prompting was deliberately excluded, since it would have added a second source of variation. The treating team's documented course is a comparator, not adjudicated truth, though it grounds the study in a real published case rather than an invented one. The case study article postdates the model's training cutoff, eliminating the possibility of training contamination.

7. Further research directions

A multi-case replication across diagnostic categories with distinct prevailing risk-factor structures would test whether the family-history anchoring effect observed here generalizes, and repeating the four-condition comparison across model versions would test whether the format-over-content pattern remains stable as models change.

8. Conclusions

In this single case, diagnostic content was materially consistent across four prompt styles; prompt style primarily affected the granularity of probability estimates and the visibility of reasoning steps. All four conditions ranked bipolar I disorder above the substance-induced etiology recorded in the treating team's documented differential, and all four recommended a broader neurologic workup than the treating team performed. These are the observations from a single vignette, offered as a proof of concept rather than as a general account of how prompt style behaves. For practicing clinicians, they suggest that prompt style is a tool for shaping how an LLM's differential is communicated and documented rather than a substitute for independent judgment about which diagnosis to weight most heavily. Testing whether that holds will require replication across multiple cases and model families, with outputs adjudicated against current clinical guidelines rather than against a single documented course.

Acknowledgments

Not applicable.

Declaration of generatiand AI-assisted technologies in the writing process

A large language model (Claude, Anthropic) was used for language editing and formatting assistance; the author reviewed and verified all content and is fully responsible for it. ChatGPT-5.5 (OpenAI) was used as the study instrument to generate the differential diagnosis outputs under evaluation, as described in Methods, and is not a use covered by this writing-process declaration.

Funding

This research received no external funding.

Institutional review board statement

Not applicable. This study involved no human participants, no new patient contact, no intervention, and no access to identifiable private information. It reanalyzed a previously published, de-identified case report [16] released under a Creative Commons Attribution license. An activity of this kind does not meet the regulatory definition of human subjects research and therefore falls outside institutional review board jurisdiction, so no approval or exemption determination was sought.

Informed consent statement

Not applicable. No new patient contact occurred. Informed consent for the original case was obtained by the treating team and reported in the source publication [16].

Data availability

The case vignette, the four prompts as submitted, and the four unedited model outputs are available on Zenodo: https://doi.org/10.5281/zenodo.21207782.

Conflict of interest

The author declares no conflicts of interest.

References

[1] Gaebe K, van der Woerd B. Evaluation of large language models as a diagnostic tool for medical learners and clinicians using advanced prompting techniques. *PLOS One*. 2025; 20(8): e0325803. doi:10.1371/journal.pone.0325803.

[2] Patil R, Heston TF, Bhuse V. Prompt engineering in healthcare. *Electronics*. 2024; 13(15): 2961. doi:10.3390/electronics13152961.

[3] Sonoda Y, Kurokawa R, Hagiwara A, Asari Y, Fukushima T, Kanzawa J, et al. Structured clinical reasoning prompt enhances LLM's diagnostic capabilities in diagnosis please quiz cases. *Japanese Journal of Radiology*. 2025; 43(4): 586-592. doi:10.1007/s11604-024-01712-2.

[4] Yao G, Zhang W, Zhu Y, Wong U, Zhang Y, Yang C, et al. Comparing the accuracy of large language models and prompt engineering in diagnosing realworld cases. *International Journal of Medical Informatics*. 2025; 203: 106026. doi:10.1016/j.ijmedinf.2025.106026.

[5] Heston TF, Khun C. Prompt engineering in medical education. *International Medical Education*. 2023; 2(3): 198-205. doi:10.3390/ime2030019.

[6] Liu J, Liu F, Wang C, Liu S. Prompt engineering in clinical practice: tutorial for clinicians. *Journal of Medical Internet Research*. 2025; 27: e72644. doi:10.2196/72644.

[7] Maaz S, Palaganas JC, Palaganas G, Bajwa M. A guide to prompt design: foundations and applications for healthcare simulationists. *Frontiers in Medicine*. 2025; 11: 1504532. doi:10.3389/fmed.2024.1504532.

[8] Sivarajkumar S, Kelley M, Samolyk-Mazzanti A, Visweswaran S, Wang Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Medical Informatics*. 2024; 12: e55318. doi:10.2196/55318.

[9] Savage T, Nayak A, Gallo R, Rangan E, Chen JH. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*. 2024; 7(1): 20. doi:10.1038/s41746-024-01010-1.

[10] Poulain R, Adiba FI, Fayyaz H, Beheshti R. Bias patterns in the application of LLMs for clinical decision support: a comprehensive study. *Delaware Journal of Public Health*. 2026; 12(1): 54-67. doi:10.32481/djph.2026.03.10.

[11] Wang J, Redelmeier DA. Artificial intelligence chain-of-thought reasoning in nuanced medical scenarios: mitigation of cognitive biases through model intransigence. *BMJ Quality & Safety*. 2026; 35(7): 456-463. doi:10.1136/bmjqs-2025-019299.

[12] Grabb D. The impact of prompt engineering in large language model performance: a psychiatric example. *Journal of Medical Artificial Intelligence*. 2023; 6: 20. doi:10.21037/jmai-23-71.

[13] Rojo-Bofill LM, Ribes Jordán G, Monedero Carrasco L, Carrasco Picazo JP, Balanzá-Martínez V, Giner L, et al. Impact of the large language model and prompt specificity in generating psychiatric multiple-choice questions. *Educación Médica*. 2026; 27(2): 101161. doi:10.1016/j.edumed.2026.101161.

[14] Sarma KV, Hanss KE, Halls AJM, Krystal A, Becker DF, Glowinski AL, et al. Integrating expert knowledge into large language models improves performance for psychiatric reasoning and diagnosis. *Psychiatry Research*. 2026; 355: 116844. doi:10.1016/j.psychres.2025.116844.

[15] Callens S. Effective prompt design for large language models in clinical practice. *Acta Clinica Belgica*. 2026; 81(2): 118-129. doi:10.1080/17843286.2026.2613903.

[16] Ho A, Liwag T, Gonzalez J, Phan A, Wong D, Fu M. Mania with psychotic symptoms in an adolescent patient associated with ashwagandha use: case report. *Open Journal of Clinical & Medical Case Reports*. 2026; 12(1): 2402.

[17] Gallifant J, Afshar M, Ameen S, Aphinyanaphongs Y, Chen S, Cacciamani G, et al. The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine*. 2025; 31(1): 60-69. doi:10.1038/s41591-024-03425-5.

[18] OpenAI. *GPT-5.5 model*. OpenAI Developers. 2026. Available from: https://developers.openai.com/api/docs/models/gpt-5.5 [Accessed 28th July 2026].

[19] Kumar A, Sah BK, Ahmad F, Jamwal V, Sahu SK. Potential adverse effects of ashwagandha: a critical review of preclinical and clinical evidence. *Phytotherapy Research*. 2026; 40(8): 4913-4931. doi:10.1002/ptr.70158.

[20] Gibson LL, Pollak TA, Blackman G, Thornton

M, Moran N, David AS. The psychiatric phenotype of anti-NMDA receptor encephalitis. *The Journal of Neuropsychiatry and Clinical Neurosciences*. 2019; 31(1): 70-79. doi:10.1176/appi.neuropsych.17120343.
[21] Brownlee S, Chalkidou K, Doust J, Elshaug AG, Glasziou P, Heath I, et al. Evidence for overuse of medical services around the world. *The Lancet*. 2017; 390(10090): 156-168. doi:10.1016/S0140-6736(16)32585-5.
[22] Ott DE. Words representing numeric probabilities in medical writing are ambiguous and misinterpreted. *JSLS: Journal of the Society of Laparoscopic & Robotic Surgeons*. 2021; 25(3): e2021.00034. doi:10.4293/JSLS.2021.00034.
[23] Jung Y, Sohn MK, Choi JY, Cho KH. A hybrid rule-based and large language model artificial intelligence system for electrodiagnostic reporting: two-center retrospective evaluation. *Muscle & Nerve*. 2026; 74(2): 390-396. doi:10.1002/mus.70280.
[24] Palm E, Manikantan A, Mahal H, Belwadi SS, Pepin ME. Assessing the quality of AI-generated clinical notes: validated evaluation of a large language model ambient scribe. *Frontiers in Artificial Intelligence*. 2025; 8: 1691499. doi:10.3389/frai.2025.1691499.
[25] Spitzer P, Hendriks D, Rudolph J, Schlaeger S, Ricke J, Kühl N, et al. The effect of medical explanations from large language models on diagnostic accuracy in radiology. *NPJ Digital Medicine*. 2026; 9(1): 333. doi:10.1038/s41746-026-02619-0.
[26] Harada Y. Decomposing persona prompts for simulated clinical reasoning: a two-by-two factorial in silico experiment of time pressure and safety prioritization. *Cureus*. 2026; 18(3): e105796. doi:10.7759/cureus.105796.
[27] Neves B, Moreira JM, Gonçalves S, Cerejo J, Da Silva NA, Leite F, et al. Zero-shot learning for clinical phenotyping: comparing LLMs and rule-based methods. *Computers in Biology and Medicine*. 2025; 192: 110181. doi:10.1016/j.compbiomed.2025.110181.
[28] Helmink FGL, Mesman E, Hillegers MHJ. Beyond the window of risk? The Dutch bipolar offspring study: 22-year follow-up. *Journal of the American Academy of Child & Adolescent Psychiatry*. 2025; 64(5): 593-601. doi:10.1016/j.jaac.2024.05.024.
[29] Martínez JA, Martínez L. Distinguishing anchoring from confirmation bias in diagnostic vignette studies: methodological implications for clinical decision-making. *Frontiers in Public Health*. 2026; 14: 1770198. doi:10.3389/fpubh.2026.1770198.
[30] Heston TF, Lewis LM. ChatGPT provides inconsistent risk-stratification of patients with atraumatic chest pain. *PLOS One*. 2024; 19(4): e0301854. doi:10.1371/journal.pone.0301854.
[31] Healy J, Kossoff J, Lee M, Hasford C. Human-AI collaboration in clinical reasoning: a UK replication and interaction analysis. *Diagnosis*. 2026. doi:10.1515/dx-2025-0176.
[32] Hopkins AM, Cornelisse E. AI can reason like a physician—what comes next? *Science*. 2026; 392(6797): 466-467. doi:10.1126/science.aeg8766.