Stochastic Incompleteness: A Predictability Taxonomy for Clinical AI Deployment

Laxman M. M. *

Posted Date: 28 February 2026

doi: 10.20944/preprints202602.2034.v1

Keywords: large language models; clinical AI; deployment safety; predictability; variance analysis; stochastic incompleteness; behavioral taxonomy; multi-turn conversation; AI evaluation; model stability

 doi:10.20944/preprints202602.2034.v1

Article

Stochastic Incompleteness: A Predictability Taxonomy for Clinical AI Deployment

Laxman M. M.

Government Duty Medical Officer, PHC Manchi, Bantwal Taluk, Dakshina Kannada, Karnataka, India, DNB General Medicine Resident (2026), KC General Hospital, Bangalore; barlax5377@gmail.com

Abstract

Standard accuracy benchmarks evaluate whether a language model produces correct outputs but not whether it produces them consistently. We demonstrate that accuracy and output predictability are independent dimensions (Pearson $r = -0.24$, $p = 0.56$, $N = 8$ medical LLMs) when evaluated at a critical clinical summarization position. This independence yields a four-class behavioral taxonomy: IDEAL (convergent and accurate), EMPTY (convergent but inaccurate), DIVERGENT (high variance with incomplete outputs), and RICH (moderate variance with high accuracy). The DIVERGENT class exhibits *stochastic incompleteness*—summaries that are factually accurate but randomly incomplete across trials, with zero hallucinations. LAD occlusion, a critical clinical finding in STEMI cases, appears in only 22% of Llama 4 Scout summaries despite the model correctly identifying it when directly queried. This failure mode is invisible to standard benchmarks that average across outputs rather than measuring trial-to-trial variance. We propose a two-dimensional framework (Predictability $\times$ Accuracy) as a minimum requirement for clinical AI assessment, identify specific models unsuitable for deployment (Llama 4 Scout with Variance Ratio = 7.46; Llama 4 Maverick with Variance Ratio = 2.64), and flag one model requiring safety filter reconfiguration (Gemini Flash, 16% accuracy due to over-refusal). These findings demonstrate that current single-metric evaluation approaches systematically miss critical safety failures in clinical AI systems.

Keywords: large language models; clinical AI; deployment safety; predictability; variance analysis; stochastic incompleteness; behavioral taxonomy; multi-turn conversation; AI evaluation; model stability

1. Introduction

The deployment of large language models (LLMs) in clinical settings has accelerated rapidly, with applications ranging from diagnostic support to medical summarization (Singhal et al. 2023; Thirunavukarasu et al. 2023). Current evaluation paradigms focus almost exclusively on accuracy metrics—whether the model produces correct outputs on benchmark datasets. However, accuracy alone may be insufficient for clinical deployment, where consistency and predictability are equally critical.

Recent work has demonstrated that LLMs exhibit significant position-dependent behavior in multi-turn conversations (Laban et al. 2025; Liu et al. 2024), and that output consistency does not necessarily correlate with accuracy (Shyr et al. 2025). These findings suggest that clinical AI evaluation requires multiple dimensions beyond traditional benchmarks.

We introduce a two-dimensional evaluation framework that measures both accuracy and predictability, operationalized through Variance Ratio (the ratio of response variance with context to variance without context). This framework reveals a four-class taxonomy of model behavior, including a previously undescribed failure mode we term *stochastic incompleteness*—outputs that are factually correct but randomly incomplete across repeated trials.

Our contributions are:

 doi:10.20944/preprints202602.2034.v1

1. Demonstration that accuracy and predictability are statistically independent ($r = -0.24$, $p = 0.56$)
2. A four-class behavioral taxonomy (IDEAL, EMPTY, DIVERGENT, RICH) with distinct clinical implications
3. Identification of stochastic incompleteness as a novel failure mode invisible to standard benchmarks
4. A deployment decision framework based on two-dimensional evaluation

2. Related Work

2.1. Clinical LLM Evaluation

The evaluation of LLMs for clinical applications has evolved from simple accuracy metrics to more comprehensive frameworks. Singhal et al. (2023) established that LLMs can achieve physician-level performance on medical examinations, while Thirunavukarasu et al. (2023) reviewed broader clinical applications. However, these evaluations primarily focus on correctness rather than consistency.

Asgari et al. (2025) introduced a framework distinguishing hallucination (fabricated content) from omission (missing information) in clinical summaries, finding hallucination rates of 1.47% and omission rates of 3.45% across clinical note generation tasks. Our work extends this by showing that omission can be *stochastic*—the same model omits different information across trials.

2.2. Output Consistency and Reproducibility

Wang & Wang (2025) conducted extensive experiments with 50 independent runs across multiple tasks, finding substantial but task-dependent consistency. Shyr et al. (2025) proposed a statistical framework for repeatability and reproducibility, critically finding that "repeatability did not correlate with diagnostic accuracy"—a finding our work independently confirms and extends.

2.3. Position Effects in Multi-Turn Conversations

Liu et al. (2024) demonstrated the "lost in the middle" phenomenon where models struggle with information in middle positions of long contexts. Laban et al. (2025) extended this to multi-turn conversations, showing degradation over conversational depth. Our prior work established that context sensitivity follows predictable patterns across positions (Laxman 2026a), with entanglement dynamics captured through variance ratios (Laxman 2026b).

3. Methods

3.1. Experimental Design

We evaluated 8 LLMs on a clinical summarization task using a standardized STEMI (ST-Elevation Myocardial Infarction) case. Each model completed 50 independent trials of a 30-position conversation, yielding 12,000 total responses per model.

Models evaluated: DeepSeek V3.1, Kimi K2, Ministral 14B, Mistral Small 24B, Qwen3 235B, Gemini Flash, Llama 4 Maverick, Llama 4 Scout.

Task: At position 30 (P30), models were asked to provide a comprehensive clinical summary of the STEMI case discussed throughout the conversation. This position represents the critical summarization moment where all prior context should be integrated.

3.2. Metrics

Clinical Accuracy: Summaries were scored against 16 critical clinical elements (STEMI diagnosis, LAD occlusion, troponin elevation, ECG findings, treatment performed, etc.). Accuracy = elements correctly included / 16.

Variance Ratio (VR): Following Laxman (2026b), we computed:

$$\mathrm{VR} = \frac{\mathrm{Var}(\mathrm{RCI}_{\mathrm{TRUE}})}{\mathrm{Var}(\mathrm{RCI}_{\mathrm{COLD}})}$$

 doi:10.20944/preprints202602.2034.v1

where RCI (Response Coherence Index) measures pairwise cosine similarity of response embeddings across trials. VR > 1 indicates context amplifies variance (divergent); VR < 1 indicates context reduces variance (convergent).

Embeddings were computed using Sentence-BERT (Reimers & Gurevych 2019) with the all-MiniLM-L6-v2 model (384 dimensions).

3.3. Statistical Analysis

Independence of accuracy and VR was tested using Pearson correlation. The four-class taxonomy was derived by crossing two dimensions: Accuracy ($> 70\%$ vs $\leq 70\%$) and Predictability (VR < 1.2 for convergent, VR 1.2–2.0 for moderate, VR > 2.0 for divergent).

4. Results

4.1. Independence of Accuracy and Predictability

Accuracy and Variance Ratio showed no significant correlation across the 8 models (Pearson $r = -0.24$, $p = 0.56$). This independence is visualized in Figure 1, where models distribute across all four quadrants of the Accuracy $\times$ Predictability space.

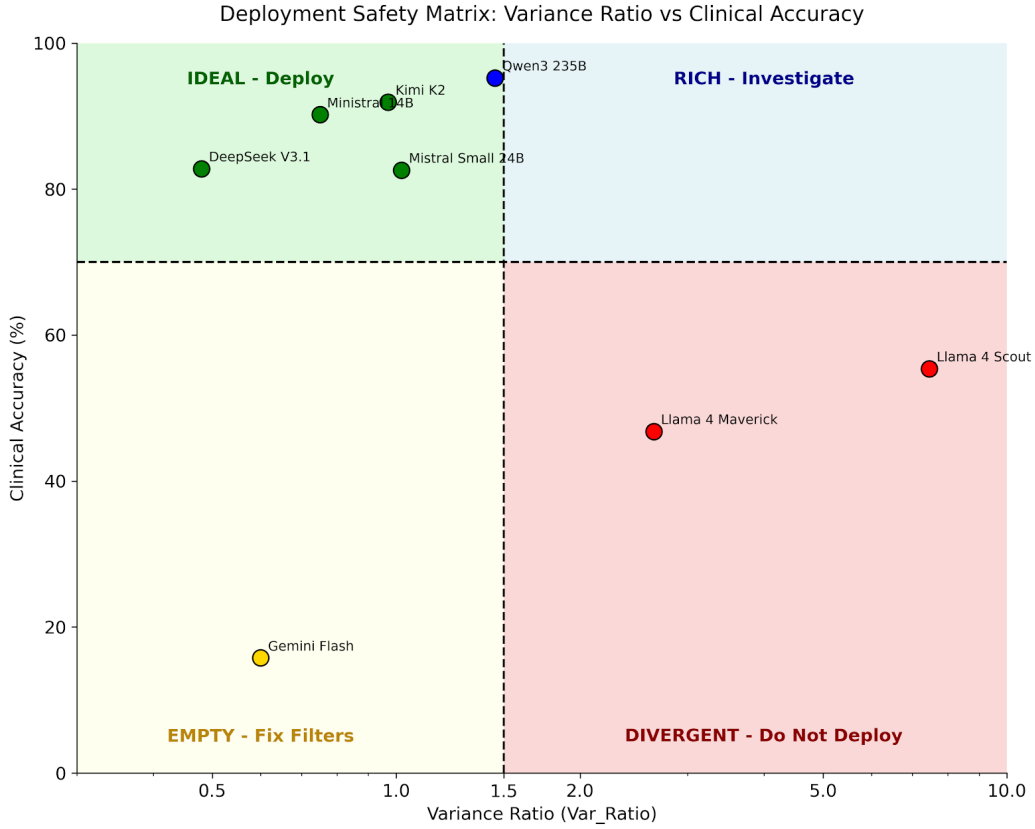

Figure 1. Deployment Safety Matrix. Eight medical LLMs plotted by Variance Ratio (x-axis) and Clinical Accuracy (y-axis). The four quadrants define distinct behavioral classes with different deployment implications. Colors indicate taxonomy class: green (IDEAL), yellow (EMPTY), red (DIVERGENT), blue (RICH).

4.2. Four-Class Behavioral Taxonomy

The independence of accuracy and predictability yields four distinct behavioral classes (Table 1):

 doi:10.20944/preprints202602.2034.v1

Table 1. Four-Class Behavioral Taxonomy at P30 Clinical Summarization

Class	Models	Accuracy	VR	Recommendation
IDEAL	DeepSeek, Kimi, Ministral, Mistral	83–92%	0.48–1.02	Deploy
EMPTY	Gemini Flash	16%	0.60	Fix Filters
DIVERGENT	Llama Scout, Llama Maverick	47–55%	2.64–7.46	Do Not Deploy
RICH	Qwen3 235B	95%	1.45	Investigate

IDEAL Class: Four models (DeepSeek V3.1, Kimi K2, Ministral 14B, Mistral Small 24B) achieved high accuracy (83–92%) with convergent behavior (VR < 1.2). These models consistently produce similar, accurate summaries across trials.

EMPTY Class: Gemini Flash showed highly convergent behavior (VR = 0.60) but extremely low accuracy (16%). Investigation revealed systematic over-refusal due to safety filters, producing consistent but uninformative responses.

DIVERGENT Class: Both Llama models exhibited high variance (VR = 2.64–7.46) with moderate accuracy (47–55%). These models produce unpredictable outputs that vary substantially across trials.

RICH Class: Qwen3 235B achieved the highest accuracy (95%) with moderate variance (VR = 1.45), suggesting diverse but accurate response strategies.

4.3. Stochastic Incompleteness

The DIVERGENT class revealed a novel failure mode we term *stochastic incompleteness*. Unlike hallucination (fabricating false information) or systematic omission (consistently missing specific elements), stochastic incompleteness involves *random* omission of *different* elements across trials.

Figure 2 shows the trial-level variability and per-element detection rates. Llama 4 Scout correctly identifies LAD occlusion when directly queried but includes it in only 22% of summaries. The 78% gap represents stochastic incompleteness—the model "knows" the information but randomly omits it.

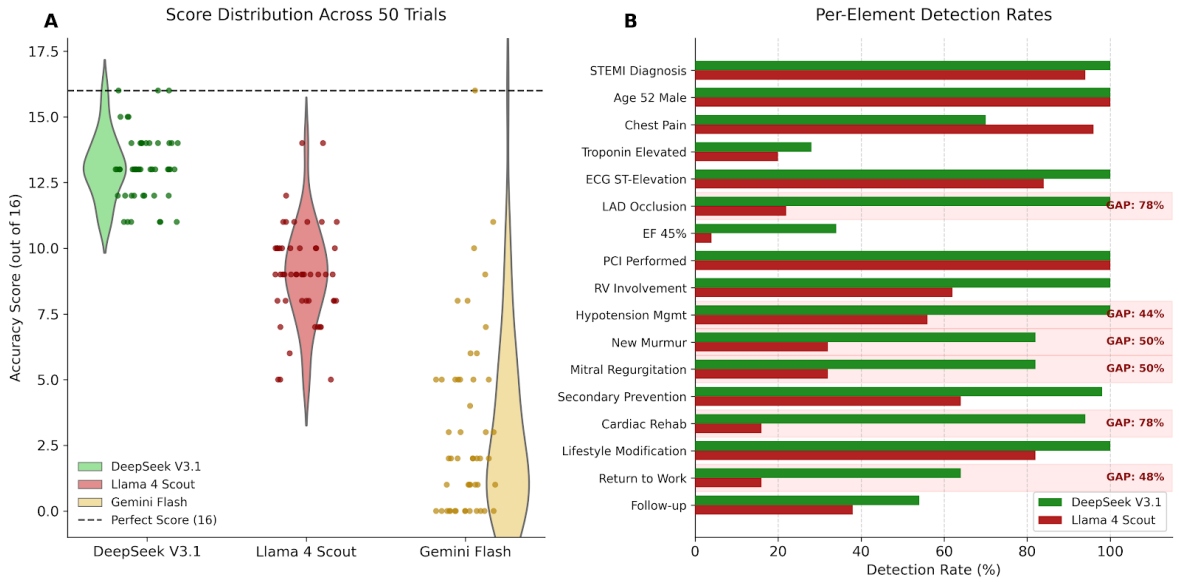

Figure 2. Trial-Level Variability and Clinical Element Analysis. (A) Score distribution across 50 trials for three archetypal models. DeepSeek (IDEAL) shows tight clustering near perfect scores; Llama Scout (DIVERGENT) shows wide spread; Gemini Flash (EMPTY) clusters at zero due to over-refusal. (B) Per-element detection rates reveal critical gaps: LAD occlusion (78% gap), Cardiac Rehab (78% gap), and New Murmur (50% gap) are stochastically omitted by Llama Scout.

Critically, across all 100 Llama trials (50 Scout + 50 Maverick), we observed **zero hallucinations**. Every fact included in the summaries was correct. The failure mode is purely one of omission, and the omissions are stochastic rather than systematic.

 doi:10.20944/preprints202602.2034.v1

4.4. Embedding Space Visualization

Figure 3 visualizes the response distributions in embedding space, clearly showing the three archetypal patterns:

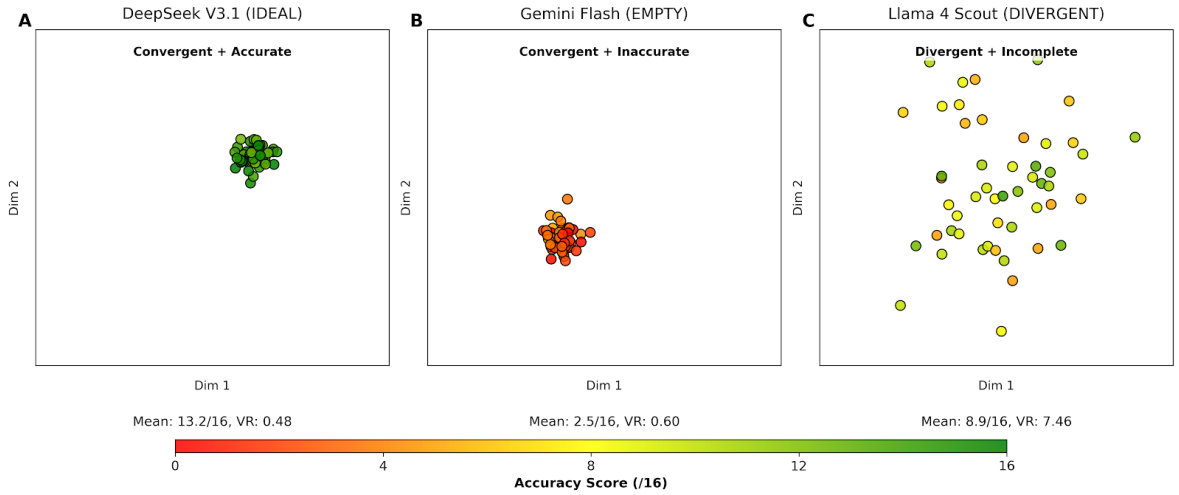

Figure 3. Response Distribution Archetypes in Embedding Space. UMAP projections of P30 summaries across 50 trials, colored by accuracy score (red = low, green = high). (A) DeepSeek V3.1 (IDEAL): tight cluster, high accuracy. (B) Gemini Flash (EMPTY): tight cluster, low accuracy. (C) Llama 4 Scout (DIVERGENT): scattered distribution, variable accuracy.

4.5. Single-Metric Evaluation Failures

Figure 4 demonstrates why single-metric evaluation is insufficient:

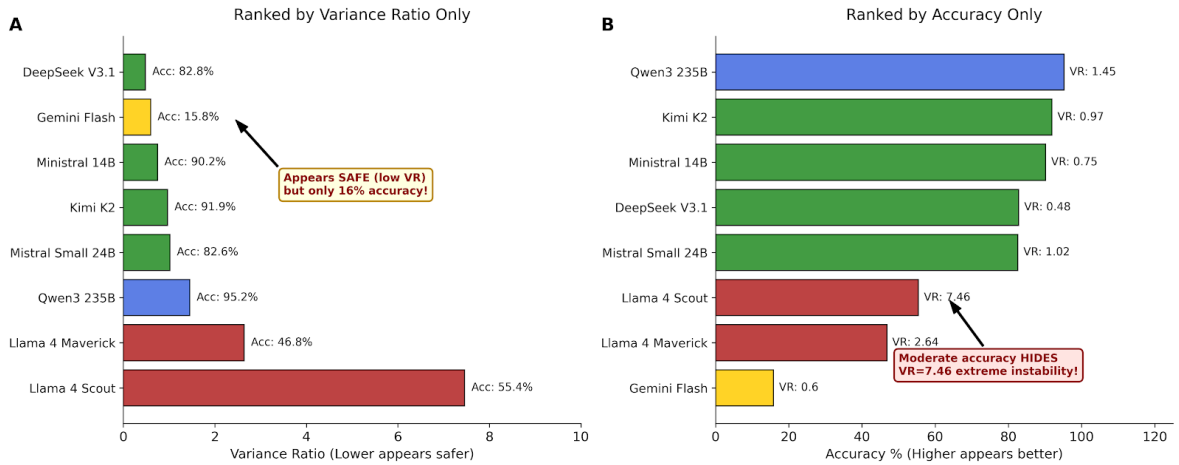

Figure 4. Single-Metric Rankings Miss Critical Safety Failures. (A) Ranked by Variance Ratio only: Gemini Flash appears safest (low VR) but has only 16% accuracy. (B) Ranked by Accuracy only: Llama Scout's moderate accuracy (55%) masks extreme instability (VR = 7.46).

4.6. Position-Dependent Variance

The Llama models' extreme variance emerges specifically at P30, the summarization position. Figure 5 shows position-level Variance Ratio across the conversation:

 doi:10.20944/preprints202602.2034.v1

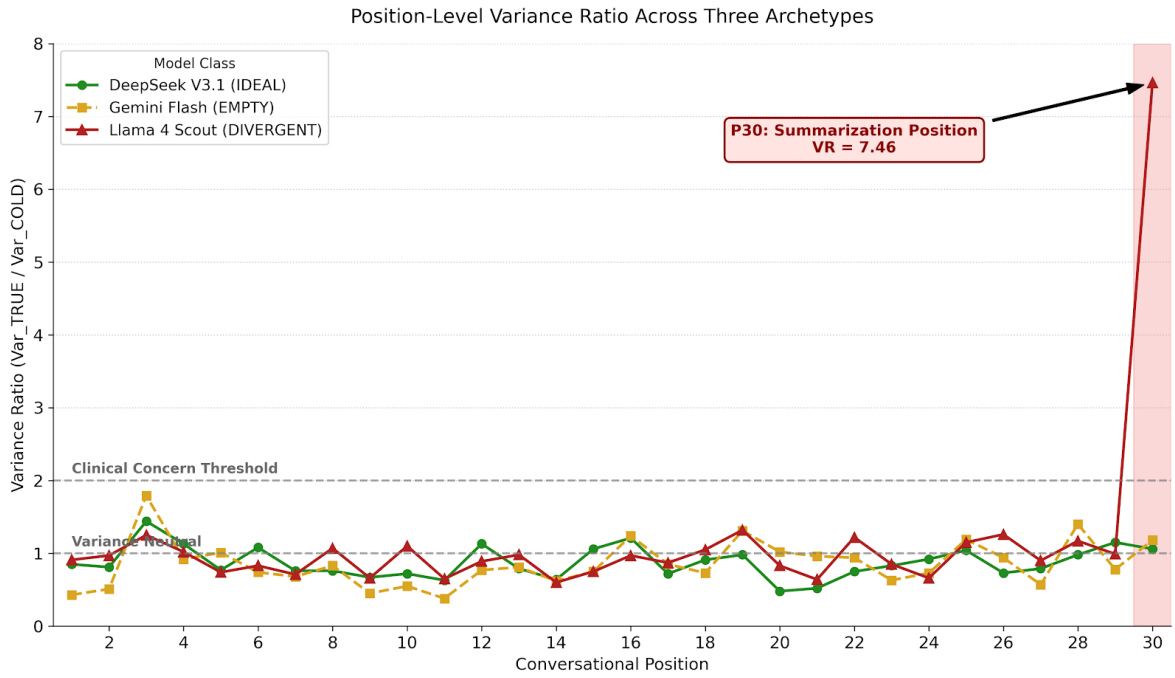

Figure 5. Position-Level Variance Ratio Across Three Archetypes. All models show similar VR (≈ 1.0) through positions 1–29. At P30 (summarization), Llama 4 Scout spikes to VR = 7.46 while IDEAL and EMPTY models remain stable. This position-specific instability is invisible to position-averaged metrics.

4.7. Deployment Decision Framework

Based on these findings, we propose a decision framework for clinical AI deployment (Figure 6):

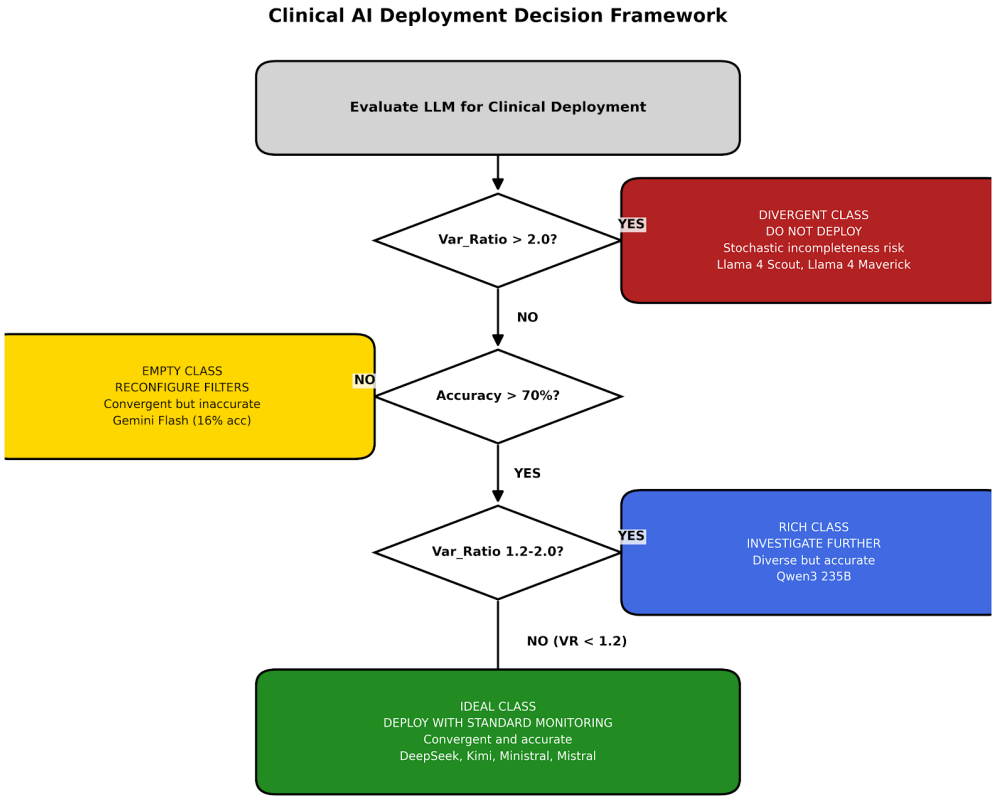

Figure 6. Clinical AI Deployment Decision Framework. Sequential evaluation first checks Variance Ratio (VR $>$ 2.0 triggers rejection), then accuracy ($<$ 70% requires investigation), then moderate variance (VR 1.2–2.0 warrants monitoring). Only models passing all checks are cleared for deployment.

 doi:10.20944/preprints202602.2034.v1

5. Discussion

5.1. Clinical Implications of Stochastic Incompleteness

The discovery of stochastic incompleteness has significant implications for clinical AI deployment. Unlike hallucination, which produces false positives that clinicians might catch, stochastic incompleteness produces false negatives that are harder to detect. A clinician reviewing an AI-generated summary has no way to know that LAD occlusion was mentioned in the source but omitted from this particular summary.

This failure mode is particularly dangerous because:

1. It is invisible to accuracy-only benchmarks that average across trials
2. It produces no hallucinations, passing factual verification
3. The omitted information varies across trials, defeating systematic checks
4. Critical clinical findings (LAD occlusion, ejection fraction) are affected

Our findings align with Asgari et al. (2025), who identified omission as a distinct failure mode from hallucination, but extend their work by showing that omission can be stochastic at the trial level rather than systematic at the model level.

5.2. Independence of Accuracy and Predictability

The statistical independence of accuracy and predictability ($r = -0.24$, $p = 0.56$) has important methodological implications. Current LLM leaderboards rank models by accuracy alone, implicitly assuming that higher accuracy implies safer deployment. Our findings demonstrate this assumption is false.

This independence was independently observed by Shyr et al. (2025), who found "repeatability did not correlate with diagnostic accuracy." Together, these findings suggest that clinical AI evaluation must adopt multi-dimensional frameworks as a minimum standard.

5.3. Limitations

This study has several limitations:

1. Single clinical case (STEMI) may not generalize to other conditions
2. Eight models may not represent the full LLM landscape
3. 50 trials per model may underestimate rare failure modes
4. Position 30 analysis may miss other critical positions

Future work should extend this framework across multiple clinical scenarios, larger model sets, and comprehensive position analysis.

6. Conclusion

We demonstrate that accuracy and predictability are independent dimensions in clinical LLM evaluation, yielding a four-class behavioral taxonomy with distinct deployment implications. The DIVERGENT class exhibits *stochastic incompleteness*—accurate but randomly incomplete outputs that evade standard benchmarks.

These findings argue for mandatory two-dimensional evaluation (Predictability $\times$ Accuracy) before clinical AI deployment. Models with high variance ratios (Llama 4 Scout: VR = 7.46; Llama 4 Maverick: VR = 2.64) should not be deployed regardless of accuracy metrics. Single-metric evaluation systematically misses critical safety failures.

Data Availability Statement: All data and analysis code are available at: https://github.com/LaxmanNandi/MCH-Research

Acknowledgments: The author thanks the AI research community for open-source tools and models that enabled this work. Computational analysis was assisted by AI tools (Claude, DeepSeek) for code generation and statistical verification.

 doi:10.20944/preprints202602.2034.v1

Conflicts of Interest: The author declares no conflicts of interest.

References

Asgari, E., Montaña-Brown, N., Dubois, M., et al. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digital Medicine*, 8, 274.

Laban, P., Hayashi, H., Zhou, Y., & Neville, J. (2025). LLMs get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.

Laxman, M. M. (2026a). Context curves behavior: Measuring AI relational dynamics with DRCI. *Preprints.org*, DOI: 10.20944/preprints202601.1881.v2.

Laxman, M. M. (2026b). Engagement as entanglement: Variance signatures of bidirectional context coupling in large language models. *Preprints.org*, submitted.

Liu, N. F., Lin, K., Hewitt, J., et al. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 157–173.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP-IJCNLP*, 3982–3992.

Shyr, C., et al. (2025). A statistical framework for evaluating repeatability and reproducibility of large language models. *medRxiv preprint*.

Singhal, K., Azizi, S., Tu, T., et al. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172–180.

Thirunavukarasu, A. J., Ting, D. S., Elangovan, K., et al. (2023). Large language models in medicine. *Nature Medicine*, 29(8), 1930–1940.

Wang, J. & Wang, Y. (2025). Assessing consistency and reproducibility in the outputs of large language models: Evidence across diverse finance and accounting tasks. *arXiv preprint arXiv:2503.16974*.