Surface Robustness in Medical Language Agents

James Smith, John Johnson, Robert Williams*

* Corresponding author: Robert Williams

Review Article | Translational Medicine and Digital Health | IASCI Research Publishing | Published 23 September 2026

ABSTRACT

How should a medical agent be tested when semantically equivalent requests produce different actions? This review answers by treating surface-form stress testing as a property of a sociotechnical workflow rather than a feature that can be read from average accuracy. The focal setting is clinical decision support under adversarial and accidental wording variation, where an error can be fluent, repeatable, and clinically consequential even when aggregate benchmark accuracy is high. Evidence from the assigned publications is synthesized with foundational studies of calibration, distribution shift, causal structure, and responsible deployment. Four requirements follow: preserve the lineage of patient signals, clinical language, retrieved evidence, attack variants, and workflow metadata; measure stability across relevant perturbations; connect confidence to a specific action; and maintain a route for human challenge and correction. The framework distinguishes descriptive performance from decision utility and separates uncertainty about the world from uncertainty created by the model and its evaluator. It also shows why faster inference or richer reasoning is valuable only when it improves a defined decision under a transparent resource budget. The article is a literature review and research agenda, not a report of a newly completed trial.

KEYWORDS

surface robustness; medical language agents; decision; medical; language; accuracy; clinical

Introduction

How should a medical agent be tested when semantically equivalent requests produce different actions? The difficulty begins with a basic observation: prediction is only one stage of a deployed process. That process selects observations, encodes them, filters competing explanations, allocates computation, and finally converts uncertainty into an action. In clinical decision support under adversarial and accidental wording variation, these choices interact with institutions and physical processes that the learned component only partially represents. A high score can coexist with fragile inputs, an evaluator that rewards the wrong surface feature, or an action rule that turns modest uncertainty into a costly decision. The resulting argument frames surface-form stress testing as an evidence problem. Its purpose is to specify the observations and comparisons needed before assurance can be claimed. For the present focus, surface-form stress testing therefore becomes a criterion for deciding which evidence deserves further scrutiny.

A process view organizes the comparison. Its unit is the complete chain from incoming signal or prompt to evidence retrieval, model reasoning, confidence, and human action. This scope makes it harder to commit a familiar error: attributing every failure to model architecture while ignoring data capture, labeling, retrieval, validation, interface design, or organizational incentives. A process-level unit also supports a more precise comparison of the assigned studies. Although their application settings differ, they each make some part of an inference chain more documented - a reward signal, a graph relation, a physical boundary condition, a confidence gate, or a revision trigger. The key test is whether that documented component remains meaningful when patient signals, clinical language, retrieved evidence, attack variants, and process metadata change, and whether the proposed safeguards lead to abstention, escalation, bounded decision support, and post-field use monitoring instead of merely producing another metric. The article's emphasis on surface-form stress testing makes that boundary observable and, in principle, auditable.

Evidence From the Focal Studies

A useful test case is Research on Security Enhancement Methods for Adversarial Robust Large Language Model Intelligent Agents for Medical Decision-Making Tasks. Rather than treating its output as an isolated score, the study frames how a medical language-model agent can combine adversarial robustness, evidence constraints, confidence control, and safe refusal. Its central mechanism is a linked pipeline for input-risk perception, evidence retrieval, knowledge-consistency checks, confidence reweighting, output control, and adversarial feedback, optimized with a multi-term objective, and its evidential basis is that the paper reports comparisons against four agent baselines and ablations under semantic perturbation, prompt injection, drug-name confusion, and false evidence. It treats safety as an end-to-end decision pathway rather than a filter attached only to the final answer. For clinical decision support under adversarial and accidental wording variation, the transferable lesson is methodological: a system should preserve the path from signal to decision and expose the conditions under which that path may fail. The boundary is equally important. The short paper and benchmark setting cannot establish clinical effectiveness, prevalence-weighted harm, or resilience to attacks outside the designed taxonomy. (Hu 2026).

Evidence for surface-form stress testing becomes concrete in DRAFT-RL: Multi-Agent Chain-of-Draft Reasoning for Reinforcement Learning-Enhanced LLMs. The paper addresses how multi-agent reinforcement learning can preserve structural diversity instead of repeatedly refining a single answer through agents generate several chain-of-draft trajectories, peer agents and a learned reward model select promising paths, and actor-critic updates feed those selections into later reasoning. Its evaluation is informative because the paper evaluates code synthesis, symbolic mathematics, and knowledge-intensive question answering and reports gains in accuracy and convergence. It treats alternative drafts as an explicit exploration space rather than incidental sampling noise. This does not make the method a universal component; it identifies a design hypothesis that must be re-tested in the article's focal setting. In particular, peer agreement can amplify correlated errors, and additional drafts increase inference and evaluation costs unless diversity is measured rather than assumed. The appropriate use of the result is therefore bounded, comparative, and explicit about unresolved deployment conditions (Li et al. 2026).

The strongest contribution of Adaptive Quantization Strategies for Robust ML Inference Under Distribution Shift is not a generic claim that learning improves performance. It is the more specific proposition that how an inference system should revise its quantization policy when deployment data no longer resembles calibration data. The proposed route is an adaptive quantization strategy that treats numerical precision as a deployment control rather than a fixed compression choice. Support comes from the fact that the study is framed around robustness under distribution shift and therefore makes the stability of low-precision inference, not compression alone, the central outcome. It connects efficient inference with the broader problem of shift-aware reliability. Read through the lens of surface-form stress testing, the study also illustrates why a reported average must remain connected to the workflow that produced it. Its conclusions should be tested across architectures, bit widths, hardware kernels, and shifts that were not represented during calibration. (Sang 2026).

Cross-Modal Generative Framework for Signal Translation from Fetal-Maternal Electrocardiograms to Fetal Doppler Waveforms asks which mechanical features of fetal Doppler waveforms can be recovered from fetal and maternal electrical signals. It uses dilated convolutions, cross-modal attention for maternal ECG, and self-attention for long-range temporal structure. The relevant evidence is that the model was trained on 885 synchronized segments from 39 pregnancies and evaluated with spectral and heart-rate reconstruction errors plus attention ablations. Separating recoverable from residual waveform components offers a computational view of electrical, maternal, and mechanical contributions. In the present review, this work matters because surface-form stress testing cannot be evaluated as an abstract virtue; it must change how evidence is collected and how an action is withheld. The cohort is small for clinical generalization, and waveform synthesis is not equivalent to validated diagnosis or improved outcomes. (Su et al. 2026).

A useful test case is Single Canonical Prompts Underestimate LLM Safety's Surface-Form Sensitivity. Rather than treating its output as an isolated score, the study frames whether one canonical wording is a faithful measurement of a model's safety behavior when intent is held constant. Its central mechanism is pre-authored, meaning-preserving reformulations applied consistently across models, human-anchored judging, intent checks, resampling controls, and paired statistical tests, and its evidential basis is that 370 seed prompts across five forms and five models show that the union of unsafe outcomes exceeds any single-form estimate and that several forms are needed to recover most observed exposure. The work treats surface form as a measurement dimension rather than a nuisance to be ignored. For clinical decision support under adversarial and accidental wording variation, the transferable lesson is methodological: a system should preserve

the path from signal to decision and expose the conditions under which that path may fail. The boundary is equally important. The form set is deliberately bounded and does not estimate a universal distribution of paraphrases, languages, or adversarial transformations. (Zhou et al. 2026).

A Layered Model of Reliability

A four-layer model exposes where assurance claims can fail. The evidence layer records observations and their provenance. The representation layer turns those observations into features, graphs, tokens, or simulated states. The inference layer produces predictions, explanations, translations, anomaly scores, or candidate actions. The decision layer maps those outputs to abstention, escalation, bounded decision support, and post-implementation monitoring. Reliability can fail at any boundary. Better inference cannot repair evidence that omitted the relevant condition, and a calibrated probability cannot rescue an action rule whose costs were never specified. Conversely, a conservative decision policy may reduce harm even when the underlying model remains imperfect. This layered view makes surface-form stress testing testable because each claim can be assigned to a component and a measurable transition. This point narrows the research task for clinical decision support under adversarial and accidental wording variation: compare alternatives under the same information and resource constraints.

Three kinds of uncertainty require different responses: aleatory variation, epistemic limits, and strategic adaptation. Aleatory variation concerns irreducible differences among cases. Epistemic uncertainty concerns missing knowledge, limited samples, or unsupported regions of the input space. Strategic adaptation occurs when users, adversaries, markets, or institutions respond to the deployed pipeline itself. These sources demand different controls. Repeated measurement can characterize variation; new evidence and conservative extrapolation can reduce epistemic uncertainty; red teaming and feedback-aware validation address adaptation. Combining them into one confidence score hides which intervention is appropriate. A defensible system therefore reports uncertainty in a form tied to an available response rather than presenting confidence as a decorative number. This point narrows the research task for clinical decision support under adversarial and accidental wording variation: compare alternatives under the same information and resource constraints.

From Evidence Capture to Escalation

The next design choice concerns where additional computation is worth its cost. Easy, well-supported cases can take a fast path. Shifted, ambiguous, or high-cost cases should activate additional precision, retrieval, alternative drafts, simulation, or expert review. This conditional design is preferable to applying maximum computation everywhere because resource cost is part of deployment quality. It is also preferable to an unconditional fast path because efficiency can amplify a systematic error at scale. The trigger must be evaluated as carefully as the expensive model: coverage, false reassurance, subgroup effects, latency, and the consequences of abstention all matter. The output is not simply a prediction but a package containing evidence links, uncertainty, the action taken, and the conditions that caused escalation. This requirement gives practical content to the article's central question: How should a medical agent be tested when semantically equivalent requests produce different actions?

The evidence architecture for clinical decision support under adversarial and accidental wording variation begins with typed evidence. Each record should identify its source, time, collection conditions, transformations, and relation to other records. Graphs are useful when relations carry meaning, but graph construction is itself a hypothesis. An edge may denote temporal adjacency, physical causation, code dependency, shared infrastructure, or analyst assertion; treating these as interchangeable creates false certainty. The architecture should therefore retain edge types and confidence, retain alternative links when evidence is ambiguous, and version both the data schema and the rules that construct the graph. A model may aggregate this structure, but the original evidence must remain recoverable for audit. For the present focus, surface-form stress testing therefore becomes a criterion for deciding which evidence deserves further scrutiny.

A Broader Evidence Base

Several established literatures clarify the design space. Selective prediction formalizes the choice to abstain when the cost of an unsupported answer exceeds the benefit of coverage (Geifman and El-Yaniv 2017). Local explanation methods remind evaluators that a plausible global description can hide unstable instance-level reasons (Ribeiro, Singh, and Guestrin 2016). Clinical language-model results demonstrate considerable knowledge while also motivating physician-centered evaluation and safety controls (Singhal et al. 2023). Responsible health-machine-learning guidance places deployment context, workflow, and monitoring beside predictive performance (Wiens et al. 2019). Together these findings imply that surface-form stress testing should be evaluated against explicit alternatives and failure costs. In Surface Robustness in Medical Language Agents, this literature supplies a specific test of surface-form stress testing within clinical decision support under adversarial and accidental wording variation.

The evidence base offers complementary tests rather than one universal metric. The clinical machine-learning literature stresses that useful systems must fit data provenance, care pathways, and prospective evaluation (Rajkomar, Dean, and Kohane 2019). Health-AI governance requires accountability, transparency, inclusion, and protection of human autonomy (World Health Organization 2021). Risk-management guidance organizes trustworthy AI as a continuing process of governance, mapping, measurement, and management (NIST 2023). They also caution against interpreting one benchmark, one split, or one explanatory artifact as a complete validation argument. In Surface Robustness in Medical Language Agents, this literature supplies a specific test of surface-form stress testing within clinical decision support under adversarial and accidental wording variation.

A credible assurance case can be built from findings that operate at different levels. Technical-debt analysis shows that data dependencies and monitoring gaps can dominate the lifecycle cost of a model (Sculley et al. 2015). Corruption benchmarks illustrate why clean-test performance is an incomplete proxy for operational robustness (Hendrycks and Dietterich 2019). Adversarial examples show that apparently small input changes can reveal large decision discontinuities (Goodfellow, Shlens, and Szegedy 2015). Their combined value lies in making assumptions visible enough to challenge before deployment and to revise after failure. In Surface Robustness in Medical Language Agents, this literature supplies a specific test of surface-form stress testing within clinical decision support under adversarial and accidental wording variation.

An Evaluation Protocol

Evaluation metrics should reflect what the system is allowed to do. Discrimination measures whether cases are ranked correctly; calibration asks whether confidence corresponds to observed frequency; selective risk asks what error remains after deferral; robustness measures performance across defined perturbations; and utility assigns costs to actions. These are not interchangeable. For surface-form stress testing, the minimum report should include overall performance, slice-level results, uncertainty intervals, coverage-risk curves, stability across runs, and failure examples linked back to patient signals, clinical language, retrieved evidence, attack variants, and workflow metadata. If the system updates incrementally, the validation must also test forgetting, stale evidence, and rollback. If an automatic judge supplies labels, a sample needs independent human review and content-preserving perturbations that reveal whether the judge is grading substance. The article's emphasis on surface-form stress testing makes that boundary observable and, in principle, auditable.

Before running a benchmark, evaluators should define a table linking claims to populations and outcomes. Each intended claim - such as improved robustness, safer abstention, faster updating, or better structural localization - needs a target population, comparator, outcome, and boundary condition. Random splits are insufficient when related samples share a patient, repository, instrument run, season, organization, or simulated design family. Grouped and temporal splits better approximate the novelty encountered after deployment. Stress tests should vary one factor at a time when attribution is important, then combine factors to measure interaction. Repeated runs are necessary whenever decoding, optimization, retrieval, or numerical kernels can vary. Reporting only the best seed or a pooled mean makes operational instability disappear. This requirement gives practical content to the article's central question: How should a medical agent be tested when semantically equivalent requests produce different actions?

Next Steps for Evidence Building

Long-term progress depends on cumulative records of both successful and failed approaches. Benchmarks should preserve negative results, failed branches, and changed definitions so later work does not learn only from successful demonstrations. Shared reporting should include data lineage, versioned evaluation code, confidence intervals, and enough error material to reproduce major conclusions. Prospective studies should predefine monitoring and stopping rules, then measure how people adapt to the application. Finally, researchers should compare simple baselines with complex architectures under equivalent information. Complexity is justified when it adds a measurable capability - for example, structural localization, calibrated deferral, or incremental updating - not merely when it creates a richer narrative around the same errors. Seen through surface-form stress testing, the issue is not merely technical; it determines which claims remain open to challenge.

A stronger evidence base requires test sets designed around operational failure modes. Cases should represent unsupported regions, ambiguous evidence, conflicting modalities, rare but costly conditions, and realistic adversarial transformations. Each case needs provenance and a reason for inclusion. The second priority is intervention testing: when uncertainty is detected, compare additional computation, retrieval, alternative reasoning paths, model ensembles, and human escalation under the same resource budget. This reveals whether a proposed safeguard actually changes the decision or only changes the explanation. The third priority is transfer. A method should be re-evaluated across sites, devices, languages, organizations, or physical regimes before broad claims are made. This point narrows the research task for clinical decision support under adversarial and accidental wording variation: compare alternatives under the same information and resource constraints.

Institutional Conditions for Reliability

Placing a person in the loop does not by itself create effective control. Reviewers need enough context to challenge the output, but not so much generated rationale that weak evidence is obscured by volume. Escalation queues require prioritization and workload limits. A reject option that sends nearly everything to experts is safe only on paper; a reject option that rarely fires may be equally ineffective. For clinical decision support under adversarial and accidental wording variation, oversight should therefore be evaluated with realistic staffing, time pressure, and disagreement. The system should record the final action and its reason without turning that record into surveillance unrelated to the stated purpose. Contestability, privacy, and security are properties of this institutional process as much as of the estimator. The article's emphasis on surface-form stress testing makes that boundary observable and, in principle, auditable.

Governance turns empirical findings into operating boundaries. A deployment specification should name allowed uses, prohibited uses, escalation thresholds, data-retention rules, and the person or role that can halt the operational tool. Monitoring should track input shift, missingness, abstention, disagreement, latency, overrides, and downstream outcomes. A rising override rate may indicate model drift, a changed operational sequence, or new user behavior; treating it only as operator noncompliance wastes evidence. Incident review should reconstruct the entire chain, including the learned component and the surrounding interface. Versioned models without versioned prompts, retrieval indexes, rules, and datasets do not support reliable reconstruction. This requirement gives practical content to the article's central question: How should a medical agent be tested when semantically equivalent requests produce different actions?

Conclusion

Surface-form stress testing is credible only when it is attached to an clearly stated evidence chain and decision policy. The review identifies several concrete mechanisms: structured exploration, typed graphs, conditional revision, adaptive computation, physical constraints, and repeated testing. None of these results makes the original benchmark a complete map of safe use. For clinical decision support under adversarial and accidental wording variation, the practical standard should be a traceable pipeline that measures shift and instability, supports abstention, escalation, bounded decision support, and post-operational use monitoring, and preserves enough evidence for an independent reviewer to reconstruct failure. Future work should compare safeguards under realistic resource and staffing limits, publish negative and subgroup results, and treat monitors and automated judges as components requiring their own validation. The final standard is not uninterrupted automation but evidence-linked action with clearly stated deferral and independent verifiability. Seen through surface-form stress testing, the issue is not merely technical; it determines which claims remain open to challenge.

References

1. Hu, Saisai. "Research on Security Enhancement Methods for Adversarial Robust Large Language Model Intelligent Agents for Medical Decision-Making Tasks." arXiv preprint arXiv:2605.08257 (2026).

2. Li, Yuanhao, et al. "DRAFT-RL: Multi-Agent Chain-of-Draft Reasoning for Reinforcement Learning-Enhanced LLMs." Proceedings of the AAAI Conference on Artificial Intelligence 40.35 (2026): 29530-29537.

3. Sang, Yinghao. "Adaptive Quantization Strategies for Robust ML Inference Under Distribution Shift." Proceedings of the 2026 5th International Conference on Cyber Security, Artificial Intelligence and Digital Economy (2026): 362-368.

4. Su, Tongli, et al. "Cross-Modal Generative Framework for Signal Translation from Fetal-Maternal Electrocardiograms to Fetal Doppler Waveforms." arXiv preprint arXiv:2607.08073 (2026).

5. Zhou, Yongxi, et al. "Single Canonical Prompts Underestimate LLM Safety's Surface-Form Sensitivity." arXiv preprint arXiv:2608.02665 (2026).

6. Geifman, Yonatan, and Ran El-Yaniv. "Selective Classification for Deep Neural Networks." Advances in Neural Information Processing Systems, vol. 30, 2017.

7. Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?: Explaining the Predictions of Any Classifier." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135-1144.

8. Singhal, Karan, et al. "Large Language Models Encode Clinical Knowledge." Nature, vol. 620, 2023, pp. 172-180.

9. Wiens, Jenna, et al. "Do No Harm: A Roadmap for Responsible Machine Learning for Health Care." Nature Medicine, vol. 25, 2019, pp. 1337-1340.

10. Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane. "Machine Learning in Medicine." New England Journal of Medicine, vol. 380, 2019, pp. 1347-1358.

11. World Health Organization. Ethics and Governance of Artificial Intelligence for Health. World Health Organization, 2021.

12. National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce, 2023.

13. Sculley, D., et al. "Hidden Technical Debt in Machine Learning Systems." Advances in Neural Information Processing Systems, vol. 28, 2015.

14. Hendrycks, Dan, and Thomas Dietterich. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations." International Conference on Learning Representations, 2019.

15. Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples." International Conference on Learning Representations, 2015.