

Designing Full-Link Security for Medical Decision Agents

Jason Green, Jeffrey Baker, Ryan Adams^{*}

^{*} Corresponding author: Ryan Adams

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

ABSTRACT

How should input screening, retrieval, reasoning, and output control be evaluated as one safety case? This review answers by treating end-to-end attack resistance as a property of a sociotechnical workflow rather than a feature that can be read from average accuracy. The focal setting is clinical question answering with external evidence, 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

designing full-link security; medical decision agents; decision; reasoning; attack; accuracy; clinical

Introduction

How should input screening, retrieval, reasoning, and output control be evaluated as one safety case? The difficulty begins with a basic observation: prediction is only one stage of a deployed operational sequence. Before an output appears, designers have already decided what counts as evidence, which representations are admissible, how alternatives are compared, and how uncertainty changes action. In clinical question answering with external evidence, these choices interact with institutions and physical processes that the predictive 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 analysis therefore approaches end-to-end attack resistance as an evidence problem. The task is therefore to identify the evidence that makes a reliability claim testable. The article's emphasis on end-to-end attack resistance makes that boundary observable and, in principle, auditable.

The argument proceeds from a operational sequence perspective. Its unit is the complete chain from incoming signal or prompt to evidence retrieval, model reasoning, confidence, and human action. The benefit is that it avoids attributing every problem to the predictor: attributing every failure to model architecture while ignoring data capture, labeling, retrieval, validation, interface design, or organizational incentives. This framing places the focal studies on comparable evidential terms. Although their application settings differ, they each make some part of an inference chain more clearly stated - a reward signal, a graph relation, a physical boundary condition, a confidence gate, or a revision trigger. The key test is whether that clearly stated component remains meaningful when patient signals, clinical language, retrieved evidence, attack variants, and operational sequence metadata change, and whether the proposed safeguards lead to abstention, escalation, bounded decision support, and post-implementation monitoring in place of merely producing another metric. For the present focus, end-to-end attack resistance therefore becomes a criterion for deciding which evidence deserves further scrutiny.

What the System Must Make Visible

The framework next distinguishes ordinary variation from missing knowledge and from actors who adapt to the system. 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 system 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 assessment 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 in place of presenting confidence as a decorative number. The article's emphasis on end-to-end attack resistance makes that boundary observable and, in principle, auditable.

Reliability becomes easier to inspect when the deployed process is divided into four layers. 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-deployment 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 end-to-end attack resistance testable because each claim can be assigned to a component and a measurable transition. The article's emphasis on end-to-end attack resistance makes that boundary observable and, in principle, auditable.

Comparative Reading of the Target Literature

Evidence for end-to-end attack resistance becomes concrete in Research on Security Enhancement Methods for Adversarial Robust Large Language Model Intelligent Agents for Medical Decision-Making Tasks. The paper addresses how a medical language-model agent can combine adversarial robustness, evidence constraints, confidence control, and safe refusal through 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. Its evaluation is informative because 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. 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, the short paper and benchmark setting cannot establish clinical effectiveness, prevalence-weighted harm, or resilience to attacks outside the designed taxonomy. The appropriate use of the result is therefore bounded, comparative, and explicit about unresolved deployment conditions (Hu 2026).

The strongest contribution of DRAFT-RL: Multi-Agent Chain-of-Draft Reasoning for Reinforcement Learning-Enhanced LLMs is not a generic claim that learning improves performance. It is the more specific proposition that how multi-agent reinforcement learning can preserve structural diversity instead of repeatedly refining a single answer. The proposed route is 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. Support comes from the fact that 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. Read through the lens of end-to-end attack resistance, the study also illustrates why a reported average must remain connected to the workflow that produced it. Peer agreement can amplify correlated errors, and additional drafts increase inference and evaluation costs unless diversity is measured rather than assumed. (Li et al. 2026).

Adaptive Quantization Strategies for Robust ML Inference Under Distribution Shift asks how an inference system should revise its quantization policy when deployment data no longer resembles calibration data. It uses an adaptive quantization strategy that treats numerical precision as a deployment control rather than a fixed compression choice. The relevant evidence is 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. In the present review, this work matters because end-to-end attack resistance cannot be evaluated as an abstract virtue; it must change how evidence is collected and how an action is withheld. Its conclusions should be tested across architectures, bit widths, hardware kernels, and shifts that were not represented during calibration.

(Sang 2026).

Connections to the Wider Literature

Several established literatures clarify the design space. Responsible health-machine-learning guidance places deployment context, workflow, and monitoring beside predictive performance (Wiens et al. 2019). 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). Together these findings imply that end-to-end attack resistance should be evaluated against explicit alternatives and failure costs. In *Designing Full-Link Security for Medical Decision Agents*, this literature supplies a specific test of end-to-end attack resistance within clinical question answering with external evidence.

The evidence base offers complementary tests rather than one universal metric. 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). They also caution against interpreting one benchmark, one split, or one explanatory artifact as a complete validation argument. In *Designing Full-Link Security for Medical Decision Agents*, this literature supplies a specific test of end-to-end attack resistance within clinical question answering with external evidence.

A credible assurance case can be built from findings that operate at different levels. Calibration research separates ranking accuracy from the trustworthiness of reported probabilities (Guo et al. 2017). Dataset-shift experiments demonstrate that uncertainty methods can deteriorate exactly when their warnings are most needed (Ovadia et al. 2019). Selective prediction formalizes the choice to abstain when the cost of an unsupported answer exceeds the benefit of coverage (Geifman and El-Yaniv 2017). Their combined value lies in making assumptions visible enough to challenge before deployment and to revise after failure. In *Designing Full-Link Security for Medical Decision Agents*, this literature supplies a specific test of end-to-end attack resistance within clinical question answering with external evidence.

Designing the Operational Workflow

A traceable operational use in clinical question answering with external evidence 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 preserve 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. Seen through end-to-end attack resistance, the issue is not merely technical; it determines which claims remain open to challenge.

Resource allocation should respond to ambiguity and consequence. 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. In clinical question answering with external evidence, this distinction should be tested before it changes an operational decision.

Evaluation That Matches the Decision

A claim-to-test matrix should precede metric selection. 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 operational use. 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 point narrows the research task for clinical question answering with external evidence: compare alternatives under the same information and resource constraints.

Metric choice must be derived from the decision. 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 end-to-end attack resistance, 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 operational sequence metadata. If the deployed operational sequence updates incrementally, the assessment 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. This requirement gives practical content to the article's central question: How should input screening, retrieval, reasoning, and output control be evaluated as one safety case?

Operational Governance and Human Review

Measurements become useful only when they alter operational use limits. A operational use specification should name allowed uses, prohibited uses, escalation thresholds, data-retention rules, and the person or role that can halt the application. Monitoring should track input shift, missingness, abstention, disagreement, latency, overrides, and downstream outcomes. A rising override rate may indicate model drift, a changed process, or new user behavior; treating it only as operator noncompliance wastes evidence. Incident review should reconstruct the entire chain, including the predictive component and the surrounding interface. Versioned models without versioned prompts, retrieval indexes, rules, and datasets do not support reliable reconstruction. This point narrows the research task for clinical question answering with external evidence: compare alternatives under the same information and resource constraints.

Oversight requires its own capacity, interface, and assessment plan. 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 question answering with external evidence, 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 learned component. This requirement gives practical content to the article's central question: How should input screening, retrieval, reasoning, and output control be evaluated as one safety case?

Research Agenda

Future testing sets should be organized around plausible mechanisms of failure. 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. The article's emphasis on end-to-end attack resistance makes that boundary observable and, in principle, auditable.

The field also needs infrastructure for cumulative, revisable evidence. Benchmarks should retain negative results, failed branches, and changed definitions so later work does not learn only from successful demonstrations. Shared reporting should include data lineage, versioned assessment 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 system. 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. The article's emphasis on end-to-end attack resistance makes that boundary observable and, in principle, auditable.

Conclusion

End-to-end attack resistance 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 assessment. The evidence also warns against treating the development benchmark as a license for unrestricted use. For clinical question answering with external evidence, the practical standard should be a traceable pipeline that measures shift and instability, supports abstention, escalation, bounded decision support, and post-deployment 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. The implication is especially consequential in clinical question answering with external evidence, where a small modeling choice can redirect attention and resources.

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. Wiens, Jenna, et al. "Do No Harm: A Roadmap for Responsible Machine Learning for Health Care." *Nature Medicine*, vol. 25, 2019, pp. 1337-1340.
5. Rajkomar, Alvin, Jeffrey Dean, and Isaac Kohane. "Machine Learning in Medicine." *New England Journal of Medicine*, vol. 380, 2019, pp. 1347-1358.
6. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health*. World Health Organization, 2021.
7. National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce, 2023.
8. Sculley, D., et al. "Hidden Technical Debt in Machine Learning Systems." *Advances in Neural Information Processing Systems*, vol. 28, 2015.
9. Hendrycks, Dan, and Thomas Dietterich. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations." *International Conference on Learning Representations*, 2019.
10. Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples." *International Conference on Learning Representations*, 2015.
11. Guo, Chuan, et al. "On Calibration of Modern Neural Networks." *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 1321-1330.
12. Ovadia, Yaniv, et al. "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift." *Advances in Neural Information Processing Systems*, vol. 32, 2019.
13. Geifman, Yonatan, and Ran El-Yaniv. "Selective Classification for Deep Neural Networks." *Advances in Neural Information Processing Systems*, vol. 30, 2017.