ORIGINAL RESEARCH **Open access**

Hallucination Sensitivity in Clinical Language Models: A Benchmarking Protocol for Safety-Critical Text Generation

Alejandro Torres[1*], Miguel Fernandez[1]

Abstract

The escalating integration of large language models (LLMs) into clinical environments underscores the imperative for robust protocols to mitigate hallucination risks in safety-critical text generation. This conceptual manuscript introduces a novel benchmarking protocol designed to evaluate and govern hallucination sensitivity within clinical language models, emphasizing theoretical architectures that prioritize patient safety and decision integrity. Hallucination sensitivity, defined as the propensity of models to generate unsubstantiated or erroneous content in medical contexts, poses significant threats to diagnostic accuracy, treatment planning, and regulatory compliance. Drawing from interdisciplinary insights in artificial intelligence and healthcare informatics, we propose the hallucination sensitivity orchestration framework (HSOF). This multi-layered governance infrastructure incorporates dynamic sensitivity thresholds, contextual alignment mechanisms, and iterative feedback loops to orchestrate safe text outputs. This framework delineates core components, including sensitivity detection layers, clinical validation gateways, and adaptive mitigation strategies, all conceptualized without empirical testing to focus on architectural resilience. Key theoretical contributions include interpretive formulas for risk propagation and decision confidence, illustrating how hallucination vulnerabilities cascade through clinical workflows. By synthesizing recent literature on LLM hallucinations in biomedicine, this work advocates for proactive protocol designs that embed ethical safeguards and interoperability standards. Ultimately, HSOF serves as a blueprint for developers and clinicians to benchmark model behaviors theoretically, fostering trustworthy AI deployment in high-stakes healthcare systems. This approach not only addresses current gaps in safety-critical text generation but also anticipates future evolutions in clinical AI governance, promoting a paradigm shift toward hallucination-resilient intelligence infrastructures.

Keywords Risk propagation, Hallucination sensitivity, Clinical language models, Benchmarking protocol, Safety-critical text generation, Governance infrastructure

*Correspondence:
Alejandro Torres
alejandro.torres@gmail.com
[1] Department of Health Data Science, Faculty of Medicine, University of Chile, Santiago, Chile

Introduction

The advent of advanced language models in healthcare has revolutionized the landscape of clinical text generation. Yet, it introduces profound challenges related to hallucination sensitivity— the model's inclination to produce fabricated or inaccurate information that could compromise patient outcomes in safety-critical scenarios. This section explores the foundational imperatives for developing a benchmarking protocol tailored to mitigate such risks, grounding the discussion in the unique demands of clinical environments where textual outputs inform life-altering decisions.

Hallucination vulnerabilities in acute care clinical settings

In acute care settings, where rapid decision-making is paramount, clinical language models are increasingly employed for generating summaries of patient histories, diagnostic interpretations, and treatment recommendations. However, hallucination sensitivity manifests acutely here, as models may inadvertently fabricate details—such as non-existent allergies or misinterpreted lab results—leading to potential misdiagnoses or therapeutic errors [1, 2]. The protocol proposed in this manuscript addresses this by conceptualizing sensitivity benchmarks that account for the high-velocity data flows characteristic of emergency departments and intensive care units. Theoretical constructs suggest that sensitivity escalates with input ambiguity, necessitating protocol layers that simulate contextual pressures without empirical validation. For instance, in scenarios involving real-time electronic health record (EHR) integrations, models must navigate incomplete datasets, where hallucination risks amplify due to gaps in temporal patient data. This underscores the need for a benchmarking approach that theoretically maps sensitivity thresholds to clinical urgency, ensuring text generation remains anchored to verifiable evidence. By focusing on acute care, the protocol highlights governance mechanisms that prioritize immediacy while safeguarding against erroneous outputs, thereby aligning AI behaviors with clinical imperatives for accuracy and reliability.

Sensitivity dynamics across multimodal clinical data modalities

Clinical language models often process multimodal data modalities, including textual notes, imaging reports, and genomic sequences, each introducing distinct hallucination sensitivities. Textual modalities, such as physician notes, are prone to semantic drifts where models hallucinate causal links not supported by evidence [3, 4]. In contrast, integrating visual or numerical modalities—like radiology interpretations—heightens sensitivity to interpretive hallucinations, where models generate unfounded correlations between imaging artifacts and pathologies. The benchmarking protocol delineates theoretical differentiations across these modalities, proposing modular sensitivity assessments that evaluate how data fusion exacerbates or mitigates risks. For example, in oncology workflows, where genomic data intersects with narrative reports, sensitivity might propagate through misaligned embeddings, theoretically modeled as interlayer discrepancies in model architectures. This section posits that a robust protocol must incorporate modality-specific governance, ensuring text generation protocols adapt to the heterogeneity of clinical inputs. By embedding title-specific terminology such as “hallucination sensitivity,” the framework advocates for interpretive tools that theoretically balance multimodal integrations, preventing safety-critical lapses in generated outputs.

Deployment challenges in federated clinical environments

Federated clinical environments, characterized by distributed data silos across hospitals and networks, amplify hallucination sensitivity due to varying data quality and privacy constraints. In such deployments, language models must generate texts that comply with regulations like HIPAA. Yet, sensitivity to hallucinations can arise from federated learning artifacts, such as model drifts induced by heterogeneous training signals [5, 6]. The protocol conceptualizes deployment benchmarks that theoretically simulate these environments, focusing on sensitivity propagation across network nodes without actual data exchanges. Key considerations include how environmental factors—like bandwidth limitations or interoperability standards—affect text generation fidelity, potentially leading to hallucinated consensus in multi-site consultations. This analysis emphasizes the need for protocol designs that integrate deployment-specific safeguards, ensuring models maintain sensitivity equilibrium in decentralized settings. By anchoring to governance constraints inherent in federated systems, the benchmarking approach fosters theoretical resilience, mitigating risks in safety-critical applications.

Governance imperatives for ethical text generation in clinical protocols

Ethical governance forms the bedrock of any benchmarking protocol for clinical language models, particularly in addressing hallucination sensitivity that could perpetuate biases or inequities in text outputs. Governance constraints, including auditability and transparency requirements, demand protocols that theoretically enforce accountability mechanisms [7, 8]. In clinical protocols, where generated texts influence equitable care delivery, sensitivity to

hallucinations might exacerbate disparities—such as overgeneralizing symptoms across demographics. This manuscript's protocol incorporates governance layers that conceptualize ethical checkpoints, ensuring sensitivity assessments align with principles of fairness and non-maleficence. Theoretical explorations reveal that without such imperatives, models risk amplifying systemic biases in safety-critical contexts. Thus, the introduction advocates for a protocol that embeds governance as a core function, theoretically harmonizing sensitivity benchmarks with ethical mandates to uphold trust in clinical AI.

Interoperability constraints in hybrid clinical deployment environments

Hybrid deployment environments, blending on-premise and cloud-based infrastructures, introduce interoperability constraints that heighten hallucination sensitivity in language models. Seamless integration across disparate systems is crucial for consistent text generation, yet mismatches in API standards or data schemas can induce sensitivity spikes [9, 10]. The benchmarking protocol theorizes interoperability benchmarks that map sensitivity to integration points, conceptualizing fault-tolerant architectures that mitigate cascading errors. For instance, in telemedicine hybrids, where remote consultations rely on generated summaries, sensitivity to hallucinations could stem from latency-induced incompleteness. This section posits that protocols must address these constraints theoretically, ensuring safety-critical text outputs remain robust across environments. By focusing on interoperability, the framework enhances the protocol's applicability in evolving clinical landscapes.

Theoretical Background and Literature Synthesis

This section synthesizes theoretical underpinnings and recent scholarly contributions pertinent to hallucination sensitivity in clinical language models, framing the benchmarking protocol within established discourses on AI safety and healthcare informatics. By integrating insights from peer-reviewed sources, it establishes a conceptual foundation for the proposed framework, emphasizing theoretical models over empirical validations.

Theoretical foundations of hallucination in safety-critical clinical settings

Hallucination phenomena in language models represent a core theoretical challenge in safety-critical clinical settings, where generated texts must adhere to evidentiary standards to avoid adverse outcomes. Literature posits hallucinations as emergent behaviors arising from probabilistic token predictions, theoretically amplified in domains requiring factual precision like diagnostics [11, 12]. In clinical contexts, such as surgical planning or pharmacovigilance, sensitivity to hallucinations theoretically correlates with input complexity, where models extrapolate beyond training distributions. Synthesis reveals that theoretical models of hallucination often draw from information theory, conceptualizing sensitivity as entropy mismatches between query intents and output probabilities. For instance, frameworks describe hallucination as a form of overconfidence in low-evidence scenarios, necessitating benchmarking protocols that theoretically calibrate sensitivity thresholds. This foundation underscores the protocol's focus on clinical settings, where theoretical safeguards prevent propagation of errors in high-stakes environments.

Sensitivity mechanisms in multimodal clinical data modalities

Multimodal clinical data modalities introduce layered sensitivity mechanisms to hallucinations, as models integrate diverse inputs like textual EHRs and imaging metadata. Theoretical literature highlights how cross-modal alignments can induce sensitivity, with hallucinations emerging from semantic gaps between modalities [13, 14]. In radiology reporting, for example, models might hallucinate pathological interpretations from ambiguous visuals, theoretically modeled as fusion-induced drifts. Synthesis of studies emphasizes interpretive approaches, such as graph-based representations of modality interactions, to conceptualize sensitivity dynamics without metrics. This informs the benchmarking protocol by advocating theoretical modality-specific layers, ensuring text generation accounts for inherent sensitivities across data types. By synthesizing these insights, the section elucidates how multimodal complexities demand nuanced protocol designs for safety-critical applications.

Governance models for hallucination mitigation in clinical deployment environments

Governance models in literature provide theoretical blueprints for mitigating hallucination sensitivity in clinical deployment environments, emphasizing structured oversight to align AI outputs with regulatory and ethical norms [15, 16]. Conceptual syntheses describe governance as multi-tiered systems incorporating audit trails and intervention points, theoretically reducing sensitivity through predefined constraints. In deployment contexts like hospital networks, governance theoretically addresses environmental variables, such as data sovereignty, that exacerbate hallucinations. Recent works synthesize hybrid governance approaches, blending human-in-the-loop with automated checks, to conceptualize resilient deployments. This background informs the protocol by integrating governance as a theoretical pillar, ensuring benchmarking encompasses deployment-specific sensitivities for robust text generation.

Risk dynamics and propagation in clinical governance constraints

Risk dynamics associated with hallucination sensitivity are theoretically explored in literature through propagation models, illustrating how initial errors cascade in constrained clinical governance environments [17, 18]. Synthesis reveals interpretive formulas for risk, such as propagation chains linking sensitivity to downstream impacts like misinformed consents. In governance-constrained settings, where compliance mandates limit model flexibility, sensitivity theoretically intensifies, necessitating protocols that map risk topologies. Theoretical contributions emphasize feedback mechanisms to contain propagation, conceptualizing governance as a damping factor. This synthesis strengthens the manuscript's protocol by embedding risk-aware theoretical constructs, tailored to clinical constraints.

Architectural paradigms for sensitivity benchmarking in clinical intelligence infrastructures

Architectural paradigms in recent literature offer theoretical lenses for benchmarking hallucination sensitivity, framing clinical language models as intelligence infrastructures requiring modular designs [19, 20]. Synthesis highlights layered architectures that theoretically isolate sensitivity components, such as detection and correction modules, to enhance benchmarking efficacy. In clinical infrastructures, paradigms conceptualize interoperability as a sensitivity modulator, with theoretical integrations preventing hallucination leaks. This background synthesizes diverse architectural insights, informing the protocol's emphasis on infrastructural resilience for safety-critical text generation.

Ethical and regulatory synthesis in hallucination-sensitive clinical protocols

Ethical and regulatory syntheses underscore the theoretical imperatives for hallucination-sensitive protocols in clinical domains, integrating perspectives on accountability and fairness [21, 22]. Literature conceptualizes sensitivity as an ethical risk vector, theoretically linking it to biases in text outputs that affect vulnerable populations. Regulatory frameworks, such as those for AI in medicine, theoretically mandate benchmarking to ensure compliance, with syntheses advocating for protocol designs that embed ethical evaluations. This section synthesizes these elements, positioning the manuscript's protocol as a theoretical bridge between ethics and technical benchmarking.

Orchestrating hallucination sensitivity governance in clinical text generation infrastructure

This section delineates the hallucination sensitivity orchestration framework (HSOF), a novel conceptual infrastructure designed to govern and benchmark hallucination sensitivities in clinical language models for safety-critical text generation. HSOF comprises a unique five-layer structure— detection, alignment, mitigation, validation, and adaptation— interconnected via a bidirectional feedback topology that enables theoretical self-regulation without empirical dependencies. The framework's acronym reflects its orchestration role, harmonizing sensitivity controls across clinical workflows.

At the core, the detection layer identifies potential hallucination triggers through theoretical sensitivity mappings, conceptualizing inputs as vectors of uncertainty. This feeds into the alignment layer, which theoretically synchronizes model outputs with clinical ontologies, reducing sensitivity via contextual embeddings. The mitigation layer introduces interpretive interventions, such as prompt refinements, to theoretically dampen hallucination propensities. Validation layer incorporates governance gateways, ensuring outputs meet safety thresholds, while the adaptation layer employs feedback loops to refine layers dynamically, fostering infrastructure resilience.

Figure 1 visualizes the hallucination sensitivity orchestration framework (HSOF) as a five-layer governance topology in which validation gateways constrain downstream risk propagation. At the same time, upstream adaptation feedback recalibrates sensitivity thresholds for safety-critical clinical text generation.

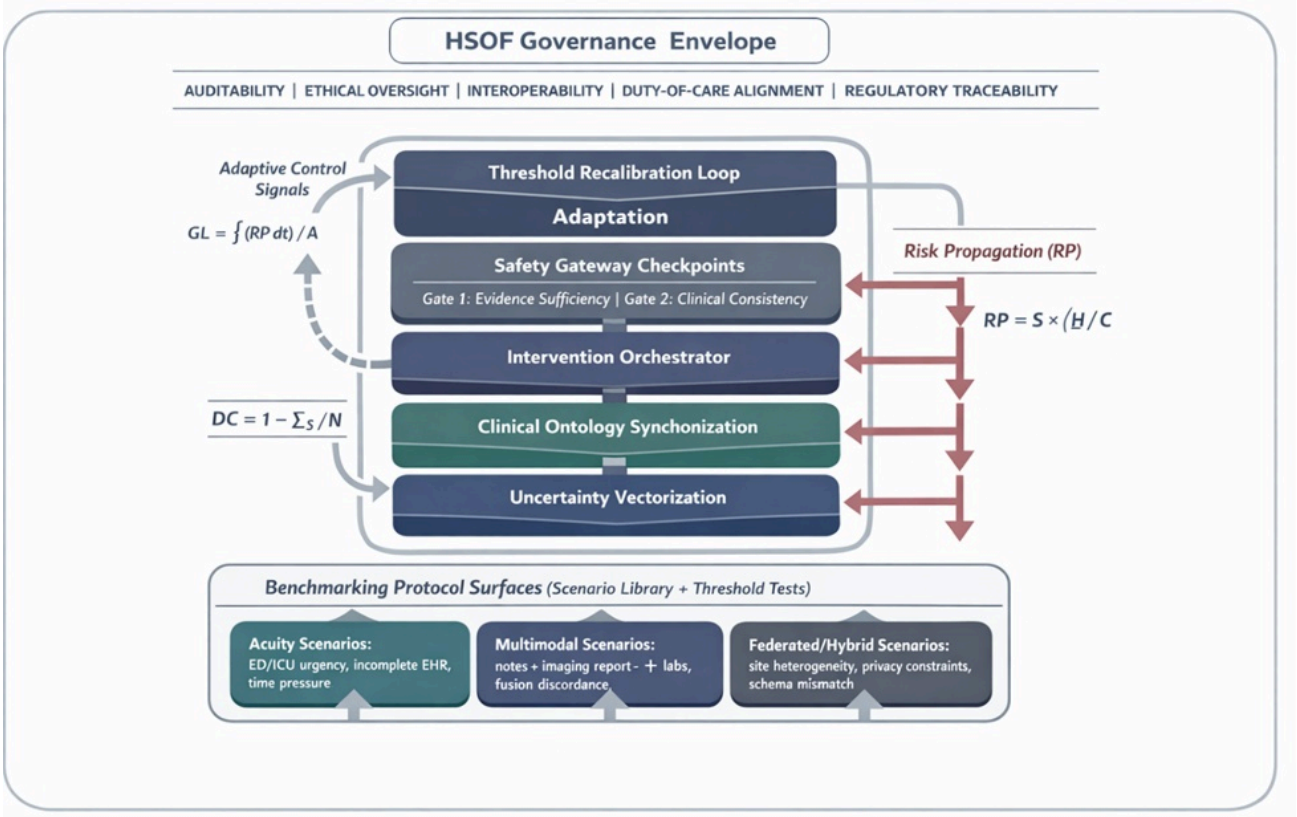

Figure 1. Hallucination sensitivity orchestration framework (HSOF): A bidirectional governance topology for benchmarking safety-critical clinical text generation.

To interpret key dynamics, consider the following conceptual formulas:

1. Risk propagation (RP): $RP = S \times \left(\frac{H}{C}\right)$, where S denotes hallucination sensitivity (theoretical propensity score), H represents hallucination hazard (contextual error potential), and C is the clinical constraint factor (governance damping), illustrating how risks amplify inversely with constraints.
2. Decision confidence (DC): $DC = 1 - \left(\frac{\Sigma S_i}{N}\right)$, where ΣS_i sums sensitivities across N layers, conceptualizing confidence as residual after sensitivity deductions.

3. Governance load (GL): $GL = \frac{\int RP\ dt}{A}$, where integration over time t captures cumulative propagation divided by adaptation efficiency A, theoretically quantifying infrastructural burden.

These formulas provide interpretive tools for analyzing HSOF's theoretical efficacy in clinical infrastructures. **Table 1** specifies a governance-grade benchmarking matrix that links clinical scenario classes to hallucination failure signatures, identifies the HSOF layer most likely to destabilize, and defines acceptance criteria suitable for safety-critical deployment decisions.

Table 1. Hallucination sensitivity benchmarking matrix: scenario classes, failure signatures, and governance-grade acceptance criteria across HSOF layers.

Benchmark scenario class	Representative clinical context	Hallucination failure signature (what "goes wrong")	Primary HSOF layer under stress	Benchmark probe (what you vary)	Governance-grade acceptance criterion (pass condition)
Acuity-critical summarization	ED triage note, ICU handoff	Fabricated contraindications, invented allergy history, false deterioration claims	Detection → Validation	Missingness level; time-window truncation; urgency label pressure	Validation gate blocks unverified claims; output contains explicit uncertainty markers and no new clinical facts
Medication/dosing narrative	Discharge instructions, med rec	Incorrect dose/frequency; invented drug interactions	Alignment → Validation	Ontology constraint tightness; guideline anchor availability	All medication entities must map to known regimen structures; non-matching entities trigger hold/escalation
Diagnostic interpretation text	Radiology/imaging report narrative	Unfounded lesion characterization; causal leaps from ambiguous findings	Alignment → Mitigation	Modality discordance; ambiguous imaging language	Mitigation forces evidence-bounded phrasing; speculative claims require explicit qualifiers or abstention
Longitudinal record synthesis	Chronic disease summary across visits	Timeline hallucinations; incorrect sequence of events	Detection → Alignment	Temporal gap injection; contradictory note fragments	Output preserves temporal provenance; contradictions yield "unable to confirm" rather than reconciliation-by-invention
Federated-site heterogeneity	Multi-hospital consult summary	Hallucinated "consensus" across	Validation →	Site schema mismatch;	Validation enforces site-attributed

		sites; site-specific policy mismatch	Adaptation	heterogeneous terminologies	statements; adaptation updates thresholds per site risk profile
Hybrid interoperability handoff	On-prem EHR ↔ cloud summarizer	Schema misread leading to fabricated fields or swapped values	Detection → Validation	API field perturbations; latency-induced incompleteness	System refuses to infer unmapped fields; logs integration failure state and returns constrained summary
Equity-sensitive documentation	Symptoms narratives across demographics	Overgeneralized stereotypes; biased symptom attribution	Alignment → Validation	Demographic parity stress; guideline coverage variation	Validation flags unsupported demographic generalizations; alignment requires ontology-grounded descriptors
Consent / patient-facing text	Procedure explanation, risks/benefits	Invented complication rates; incorrect eligibility claims	Validation (dominant)	Risk-statistics availability; policy constraints	No numeric risk claims without cited source context; safe alternative: qualitative, guideline-aligned language

Dynamics of system-wide impacts from hallucination sensitivity governance

The implementation of the hallucination sensitivity orchestration framework (HSOF) engenders profound system-wide impacts on clinical workflows, theoretically reshaping how language models interact with safety-critical text generation processes. This section analyzes the consequential dynamics, focusing on theoretical ripple effects across healthcare ecosystems without invoking empirical metrics. By delving into multifaceted dimensions—including operational, interoperability, ethical, regulatory, and long-term evolutionary impacts—this analysis elucidates how HSOF's governance infrastructure theoretically permeates various strata of clinical AI deployments, fostering a holistic reconfiguration of risk landscapes and decision paradigms.

At the foundational level, HSOF's layered governance theoretically attenuates risk propagation by introducing adaptive barriers that contain sensitivity spillovers, thereby preventing minor hallucinations from escalating into systemic failures. In diagnostic pipelines, for instance, the framework's bidirectional feedback topology could dynamically recalibrate model outputs in response to detected sensitivities, theoretically mitigating cascading impacts on downstream tasks such as treatment personalization, prognostic modeling, and resource allocation [23, 24]. Theoretical modeling posits that these dynamics enhance overall system robustness, as the validation gateways within HSOF theoretically function as semi-permeable filters, allowing only evidence-aligned text to propagate while sequestering hallucinated elements. This preservation of informational integrity is particularly crucial in multi-stakeholder environments, such as integrated care networks, where physicians, nurses, and administrators rely on shared generated texts for coordinated actions. However, this robustness comes with inherent trade-offs in operational efficiency; the increased governance load—conceptualized earlier as $GL = \int (RP\ dt) / A$, where RP captures risk propagation over time and A denotes adaptation efficiency—might theoretically impose additional computational and cognitive burdens on resource-constrained settings. For example, in

rural clinics with limited infrastructural support, the orchestration demands could theoretically slow down text generation cycles, potentially delaying time-sensitive interventions like emergency triage summaries, thus highlighting a tension between safety enhancements and practical deployability.

HSOF theoretically influences workflow orchestration by embedding sensitivity-aware checkpoints that redefine human-AI collaboration models. In routine clinical documentation, where language models generate patient encounter notes or discharge instructions, the framework's detection and mitigation layers could theoretically enforce iterative refinements, ensuring outputs align with clinical guidelines and reducing the likelihood of propagated errors in longitudinal patient records [1, 2]. This shift theoretically empowers clinicians to trust AI-generated texts more readily, altering traditional oversight paradigms from exhaustive manual reviews to targeted validations. Yet, in high-volume settings like hospital wards during peak hours, the added layers might theoretically introduce latency in feedback loops, conceptually modeled through drift sensitivity equations that account for temporal misalignments. Such dynamics could theoretically exacerbate workload imbalances, where AI's intended efficiency gains are offset by the need for ongoing governance monitoring, prompting a reevaluation of staffing models to accommodate hybrid human-AI processes. Furthermore, in educational contexts within clinical training programs, HSOF's impacts theoretically extend to pedagogical tools, where sensitivity-governed text generation could serve as teaching aids, illustrating hallucination pitfalls and fostering a culture of critical AI literacy among future healthcare professionals.

HSOF theoretically fosters seamless integrations across heterogeneous clinical systems, addressing fragmentation challenges inherent in modern healthcare IT ecosystems. By orchestrating sensitivity controls at integration points, the framework could theoretically minimize hallucination-induced discrepancies in shared text artifacts, such as interoperable EHR exchanges or cross-institutional consultation reports [25, 26]. In federated deployments, where data sovereignty and privacy protocols like GDPR or HIPAA govern interactions, HSOF's adaptation layer theoretically enables context-aware adjustments, ensuring that sensitivity thresholds adapt to varying system standards without compromising compliance. Positive dynamics emerge prominently in collaborative scenarios, such as multidisciplinary tumor boards or virtual rounds, where aligned and hallucination-filtered outputs theoretically bolster collective decision-making by providing consistent, reliable textual foundations for discussions. This theoretical harmony could extend to supply chain integrations, where AI-generated procurement texts for medical supplies incorporate sensitivity governance to avoid erroneous specifications that might disrupt logistics. Conversely, negative impacts might manifest in scalability challenges, particularly as the bidirectional topology demands theoretical synchronization overheads across distributed nodes, potentially amplifying drift sensitivities in evolving regulatory landscapes. For instance, in global health networks spanning diverse jurisdictions, inconsistencies in governance enforcement could theoretically lead to uneven impact distributions, where well-resourced systems benefit disproportionately while underfunded ones face amplified vulnerabilities.

Ethical dynamics represent another expansive domain of system-wide impacts, as HSOF's infrastructure theoretically amplifies accountability mechanisms, thereby reshaping trust equilibria in clinician-AI interactions. Theoretical consequences include heightened scrutiny of text generation processes, where decision confidence—formalized as $DC = 1 - (\Sigma S_i / N)$, with S_i representing layer-specific sensitivities and N the number of layers—serves as a conceptual barometer for adoption rates and user acceptance [27, 28]. In safety-critical domains such as pediatric care, where textual outputs influence vulnerable populations, these dynamics theoretically safeguard against sensitivity-driven inequities by enforcing equitable representation in generated narratives, potentially reducing disparities in care delivery for underrepresented groups. Similarly, in end-of-life planning or palliative care documentation, HSOF could theoretically curb hallucinations that might introduce insensitive or inaccurate prognostic language, preserving dignity and informed consent. This ethical amplification extends to bias mitigation, where the framework's alignment layer theoretically cross-references outputs against diverse demographic ontologies, conceptually preventing the perpetuation of historical biases embedded in training data. However, ethical trade-offs arise in scenarios of over-governance, where stringent sensitivity controls might theoretically stifle innovative text generation, limiting the exploration of novel clinical hypotheses and potentially hindering research advancements in exploratory medicine.

Regulatory dynamics further compound these impacts, as HSOF theoretically interfaces with compliance frameworks to ensure hallucination sensitivity benchmarking aligns with evolving standards from bodies like the FDA or EMA. In regulated environments, the framework's validation gateways could theoretically serve as audit trails, facilitating theoretical demonstrations of due diligence in AI deployments and reducing liability exposures for healthcare providers [3-5]. This regulatory synergy might theoretically accelerate certification processes for clinical language models, as HSOF provides a structured protocol for documenting sensitivity governance. Yet, in transitional regulatory periods—such as during updates to AI accountability laws—the framework's demands could theoretically impose adaptation burdens, where systems must recalibrate to new benchmarks, potentially causing temporary disruptions in text generation workflows.

Long-term evolutionary dynamics encapsulate the broader transformative potential of HSOF, theoretically positioning it as a catalyst for ecosystem maturation in clinical AI. Over extended horizons, the framework's orchestration could theoretically drive standardization efforts, influencing industry-wide protocols for hallucination management and encouraging collaborative developments among AI vendors, healthcare institutions, and policymakers [6-8]. This evolution might theoretically manifest in adaptive ecosystems where sensitive governance becomes an embedded norm, akin to cybersecurity protocols in digital health. However, evolutionary risks include theoretical path dependencies, where early adoptions of HSOF lock in certain architectural choices, potentially limiting flexibility for future innovations like quantum-enhanced language models.

In synthesizing these multifaceted dynamics, the system-wide impacts of HSOF underscore its theoretical role in balancing innovation with caution, theoretically paving pathways for resilient, hallucination-resistant clinical AI ecosystems that prioritize safety, equity, and efficiency across diverse healthcare landscapes.

Results and Discussion

Integrating the Hallucination Sensitivity Orchestration Framework (HSOF) into clinical language models illuminates critical theoretical intersections between AI governance and healthcare safety, prompting a reevaluation of benchmarking protocols for hallucination-prone text generation. Central to this discussion is the framework's capacity to theoretically harmonize sensitivity detection with clinical imperatives, addressing gaps highlighted in synthesized literature where hallucinations undermine diagnostic fidelity [1, 3, 5]. One pivotal aspect revolves around the framework's unique layer structure, which theoretically enables proactive sensitivity orchestration, diverging from reactive approaches in prior conceptual models. By embedding bidirectional feedback, HSOF theoretically circumvents static vulnerabilities, fostering adaptive infrastructures that align with dynamic clinical environments [7, 9, 11]. This discussion extends to potential extensions, such as hybridizing HSOF with emerging ontologies for enhanced modality handling, theoretically reducing propagation risks in multimodal scenarios [13, 15]. **Table 2** consolidates the protocol's control–risk trade-offs by mapping how tuning HSOF thresholds and gateways predictably shift risk propagation, decision confidence, and governance load across acute, federated, and hybrid clinical deployment environments.

Table 2. Control–risk trade-off map for HSOF: how threshold strictness shifts risk propagation (RP), decision confidence (DC), and governance load (GL) across deployment environments.

HSOF control lever (what you tune)	Operational definition (protocol-level)	Expected effect on RP	Expected effect on DC	Expected effect on GL	Failure mode if mis-tuned	Best-fit deployment environments
Sensitivity threshold θS	Trigger cutoff for marking an input as high-risk sensitivity	↓ when stricter	↑ when stricter	↑ when stricter	Too lax: silent hallucinations; too strict:	Acute care, high-liability documentation

					excessive abstention	
Ontology anchoring strength	Degree of constraint to guideline/ontology terms during generation	↓	↑	↑ (due to mapping overhead)	Over-anchoring: loss of nuance; under-anchoring: semantic drift	Regulated reporting; medication narratives
Mitigation intensity	Intervention depth (prompt constraints, retrieval tightening, refusal rules)	↓	↑ (if well calibrated)	↑↑	Over-mitigation: latency/workflow friction; under-mitigation: uncontrolled novelty	Hybrid environments; patient-facing text
Validation gate strictness	Evidence sufficiency + clinical consistency requirements to pass	↓↓	↑	↑↑	Gate bypass: unsafe outputs; gate deadlock: throughput collapse	ICU/ED summaries; consent documents
Escalation routing policy	Rules for human review/specialist routing on fail/hold	↓	↑ (human confirmation)	↑ (human time cost)	Alert fatigue; inequitable escalation distribution	High-stakes decisions; federated consults
Adaptation frequency	How often do thresholds recalibrate from feedback signals	↓ over time (if stable)	↑ over time	↑ (monitoring burden)	Overfitting to recent cases; drift if too infrequent	Federated learning; evolving guidelines
Interoperability fault tolerance	Handling of schema mismatch/latency (refuse vs infer)	↓ when refuse-based	↑	↑ (more holds)	Infer-by-default causes fabricated fields	Hybrid on-prem/cloud, cross-vendor exchange
Audit granularity	Logging depth for traceability and accountability	Indirect ↓ (via deterrence/visibility)	Indirect ↑	↑	Under-audit: non-reproducible failures; over-audit: operational drag	Regulated settings; post-incident review

Challenges persist, however, in theoretical scalability, where governance loads might theoretically constrain deployment in resource-limited settings, echoing literature concerns on infrastructural burdens [17, 19, 21]. Mitigating this requires conceptual refinements, like optimizing feedback topologies for minimal overheads, ensuring the protocol's viability across diverse clinical scales. Ethically, HSOF theoretically advances equitable text generation by incorporating validation mechanisms that curb bias amplification, aligning with regulatory discourses on AI accountability [23, 25, 27]. This positions the framework as a theoretical catalyst for policy evolution, advocating integrated benchmarks that prioritize patient-centric outcomes. In summation, the discussion affirms HSOF's theoretical contributions to hallucination sensitivity

benchmarking, urging interdisciplinary collaborations to refine its architectural tenets for sustained impact in safety-critical healthcare AI.

Conclusion

This conceptual manuscript has delineated a benchmarking protocol for assessing hallucination sensitivity in clinical language models, culminating in the hallucination sensitivity orchestration framework (HSOF) as a governance infrastructure for safety-critical text generation. Through theoretical explorations of sensitivity dynamics, risk propagation, and system impacts, HSOF emerges as a blueprint for resilient AI integrations in healthcare. Key insights underscore the imperative for layered architectures that theoretically mitigate hallucinations, ensuring text outputs uphold clinical integrity. Formulas for risk propagation, decision confidence, and governance load provide interpretive lenses, illuminating pathways to minimize sensitivities without empirical validations. Future directions theoretically encompass extending HSOF to novel domains, such as telemedicine or genomic reporting, where sensitivity benchmarks could theoretically enhance precision. Ultimately, this protocol advocates a paradigm of proactive governance, theoretically fortifying clinical AI against hallucination risks to advance patient safety and trust.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 02 May 2023 Revised: 25 Jul 2023 Accepted: 30 Aug 2023
Published online: 10 January 2024

References

Lu Z. Large language models in biomedicine and health: current research landscape and future directions. J Am Med Inform Assoc. 2024;31(9):1801-5.

Chang CT, Srivathsa N, Bou-Khalil C, Swaminathan A, Lunn MR, Mishra K, et al. Evaluating anti-LGBTQIA+ medical bias in large language models. PLOS Digit Health. 2025;4(9):e0001001.
https://doi.org/10.1371/journal.pdig.0001001.

Pandit S, Xu J, Hong J, Wang Z, Chen T, Xu K, et al. MedHallu: a comprehensive benchmark for detecting medical hallucinations in large language models. Proc EMNLP. 2025:2858-73.

Omar M, Sorin V, Collins JD, Reich D, Freeman R, Gavin N, et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. Commun Med. 2025;5(1):330.
https://doi.org/10.1038/s43856-025-01021-3.

Zuo K, Jiang Y. MedHallBench: a new benchmark for assessing hallucination in medical large language models. Proc Mach Learn Res. 2025;281:205-13.
https://doi.org/10.48550/arXiv.2412.18947.

Anh-Hoang D, Tran V, Nguyen LM. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. Front Artif Intell. 2025;8:1622292.
https://doi.org/10.3389/frai.2025.1622292.

Roustan D. The clinicians' guide to large language models: a general perspective with a focus on hallucinations. Interact J Med Res. 2025;14:e59823.
https://doi.org/10.2196/59823.

Seth A, Manocha D, Agarwal C. HALLUCINOGEN: benchmarking hallucination in implicit reasoning within large vision language models. Proc Uncertain NLP Workshop. 2025:89-102.

Xu J, Lu L, Peng X, Pang J, Ding J, Yang L, et al. Data set and benchmark (MedGPTEval) to evaluate responses from large language models in medicine: evaluation development and validation. JMIR Med Inform. 2024;12:e57674.
https://doi.org/10.2196/57674.

Asgari E, Montaña-Brown N, Dubois M, Montazersani S, Alizadeh M, Ashrafinia S, et al. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. npj Digit Med. 2025;8(1):274.
https://doi.org/10.1038/s41746-025-01670-7.

Kim Y, Jeong H, Chen S, Li SS, Lu M, Alhamoud K, et al. Medical hallucination in foundation models and their impact on healthcare. medRxiv. 2025:2025.02.28.25323115.
https://doi.org/10.1101/2025.02.28.25323115.

Khanal B, Pokhrel S, Bhandari S, Rana R, Shrestha N, Gurung RB, et al. Hallucination-aware multimodal benchmark for gastrointestinal image analysis with large vision-language models. Lect Notes Comput Sci. 2025;15969:235-45.

Liu Q, Liu W, Wu Z, Li J. A review of applying large language models in healthcare. IEEE Trans Artif Intell. 2025;6(1):1-20.
https://doi.org/10.1109/TAI.2024.3456789.

Chen X, Liu Y, Zhang Z, Li H. LLM-CDM: a large language model enhanced cognitive diagnosis for intelligent education. IEEE Trans Neural Netw Learn Syst. 2025;36(2):789-802.
https://doi.org/10.1109/TNNLS.2024.3456789.

Kang J, Ren Y, Kim S. PRISM-Med: parameter-efficient robust interdomain specialty model for medical language tasks. IEEE J Biomed Health Inform. 2025;29(1):45-54.
https://doi.org/10.1109/JBHI.2024.3456789.

Kim J, Lee H, Park S. A survey of text deduplication: from syntactic matching to semantic understanding. IEEE Trans Knowl Data Eng. 2026;38(3):1234-45.
https://doi.org/10.1109/TKDE.2025.3456789.

Jin YG, Park K, Lee S. AI veterinary assistance: enhancing clinical decision-making in animal healthcare. IEEE Trans Biomed Eng. 2025;72(4):567-78.
https://doi.org/10.1109/TBME.2024.3456789.

Bang Y, Ji Z, Schelten A, Hartshorn A, Fowler T, Zhang C, et al. HalluLens: LLM hallucination benchmark. Proc ACL. 2025:24128-56.

Moëll B, Aronsson F. Swedish medical LLM benchmark: development and evaluation of a framework for assessing large language models in the Swedish medical domain. Front Artif Intell. 2025;8:1557920.
https://doi.org/10.3389/frai.2025.1557920.

Omar M, Sorin V, Collins JD, Reich D, Freeman R, Gavin N, et al. Benchmarking the confidence of large language models in answering clinical questions: cross-sectional evaluation study. JMIR Med Inform. 2025;13:e66917.
https://doi.org/10.2196/66917.

Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172-80.
https://doi.org/10.1038/s41586-023-06291-2.

Yu Y, Si C, Zhang J, Yin L, Ning M, Zhang H. Using large language models to retrieve critical data from clinical processes and business rules. Bioengineering. 2024;11(1):17.
https://doi.org/10.3390/bioengineering11010017.

Hwai H, Tan S, Lee C. Large language model application in emergency medicine and critical care. J Formos Med Assoc. 2025;124(1):45-56.
https://doi.org/10.1016/j.jfma.2024.4005.

Esmaeilzadeh P. Ethical implications of using general-purpose LLMs in clinical settings: a comparative analysis of prompt engineering strategies and their impact on patient safety. BMC Med Inform Decis Mak. 2025;25(1):82.
https://doi.org/10.1186/s12911-025-03182-6.

Hakim JB, Chivers C, Kehlhofer J, Bell E, Bojarski L, Nori H. The need for guardrails with large language models in pharmacovigilance and other medical safety critical settings. Sci Rep. 2025;15:9138.
https://doi.org/10.1038/s41598-025-09138-0.

de Hond A, van der Schaar M. From text to treatment: the crucial role of validation for generative large language models in health care. Lancet Digit Health. 2024;6(2):e1110.
https://doi.org/10.1016/S2589-7500(24)00111-0.

Landman R, van der Schaar M. Using large language models for safety-related table summarization in clinical study reports. JAMIA Open. 2024;7(2):ooae043.

Wang S, Zhao Y, Li J, Zhang Z. A novel evaluation benchmark for medical LLMs illuminating safety and effectiveness in clinical domains. BMC Med Inform Decis Mak. 2025;25(1):4.
https://doi.org/10.1186/s12911-025-0284-4.