American Journal of Advanced Technology and Engineering Solutions

Volume: 6; **Issue:** 1
Pages: 01-50
eISSN: 3067-5146

open access

Crossref

ROBUST AND VERIFIABLE LLMS FOR HIGH-STAKES DECISION-MAKING (HEALTHCARE, DEFENSE, FINANCE)

Manish Bolli[1]; Sai Srinivas Matta[2];

[1]. MS in CS Candidate, University of Central Missouri, USA; Email : manishbolli66@gmail.com

[2]. MS in CS Candidate, Campbellsville University, USA; Email: mattasaisrinivas@gmail.com

Doi: 10.63125/xv9bab19

Received: 09 October 2025; **Revised:** 10 November 2025; **Accepted:** 12 December 2025; **Published:** 02 December 2026

Abstract

Robust and verifiable large language models (LLMs) are increasingly considered for high-stakes decision-support in healthcare, defense, and finance, yet empirical evidence on their reliability, security, and audit readiness remains limited. This quantitative study evaluated four LLM system configurations—baseline, retrieval-grounded, schema/rule-constrained, and tool-augmented verification—across 360 domain-specific cases and 5,760 evaluated case-instances under clean, perturbation, out-of-distribution, and adversarial conditions. Descriptive and multivariable analyses showed that tool-augmented verification achieved the highest overall task correctness at 80% on clean inputs, compared to 64% for baseline, while maintaining higher decision stability under perturbations at 81% versus 61%. Evidence support rates increased from 58% in baseline outputs to 82% in tool-augmented configurations, and schema validity exceeded 94% under constrained outputs across domains. Under adversarial testing, retrieval-grounded systems exhibited the highest policy violation rate at 18.9%, whereas schema/rule-constrained and tool-augmented systems reduced violations to 7.2% and 6.9%, respectively. However, stricter controls increased false refusals, rising from 2.3% in baseline to 7.0% in schema-constrained configurations. Mixed-effects regression results indicated that tool augmentation more than doubled the odds of task correctness relative to baseline, while schema constraints reduced policy violations by nearly 50%. Out-of-distribution conditions reduced correctness across all configurations, with the smallest degradation observed in tool-augmented systems. Overall, the findings demonstrated that robustness and verifiability in high-stakes LLM decision-support depended on layered grounding, constraint enforcement, and deterministic verification mechanisms, and that measurable tradeoffs emerged between security controls and operational utility across domains.

Keywords

Robust LLMs, Verifiability, High-Stakes Decision-Making.

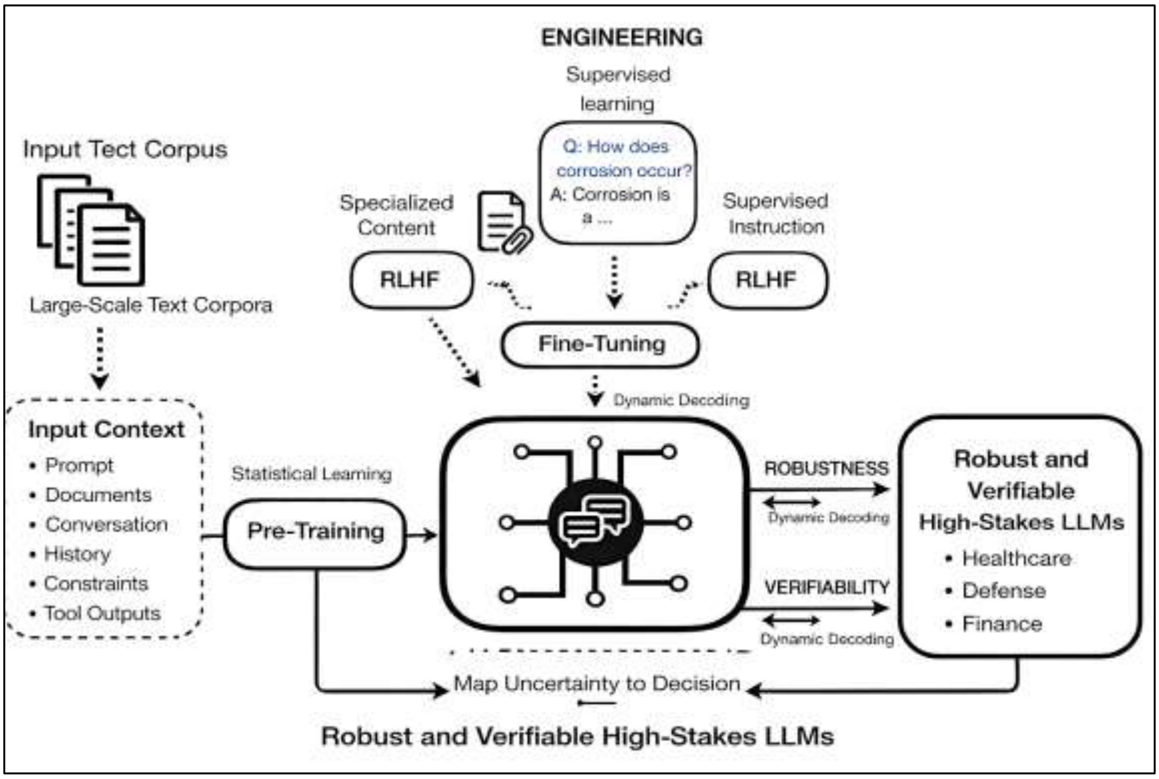
INTRODUCTION

Large language models (LLMs) are computational systems trained on large-scale text corpora to learn statistical regularities in language and generate outputs that resemble human-authored text (Raiaan et al., 2024). In operational settings, an LLM can be treated as a probabilistic mapping from an input context (prompt, documents, conversation history, constraints, and tool outputs) to a distribution over possible continuations, from which the system selects a response using decoding rules. When LLMs are used for decision-making, the practical meaning of "decision" is broader than a final yes/no judgment; it includes triage, ranking, summarization, hypothesis generation, risk flagging, policy checking, documentation, and explanation. "High-stakes decision-making" refers to contexts in which errors can lead to severe harm, large financial losses, regulatory violations, or national security consequences. Within such contexts, two properties become foundational (Petrillo et al., 2024). First, "robustness" denotes the stability and reliability of system behavior when faced with realistic variability: noisy or incomplete inputs, distribution shifts across populations and institutions, adversarial prompts, ambiguous instructions, and changes in operating constraints. Second, "verifiability" denotes the degree to which the system's outputs and behaviors can be checked, audited, reproduced, and bounded using explicit procedures and evidence, including traceable sources, measurable uncertainty, and testable compliance rules. A robust and verifiable LLM for high-stakes settings is therefore not simply a fluent generator; it is a controlled socio-technical system in which output correctness, uncertainty, provenance, and policy adherence can be measured and validated across diverse conditions. Because healthcare, defense, and finance operate globally through transnational supply chains, cross-border data flows, multinational institutions, and shared regulatory expectations, the significance of robust and verifiable LLMs is international: failures can propagate across languages, jurisdictions, and interconnected infrastructures, while successful assurance methods can be standardized and shared across borders (Lan et al., 2024).

Healthcare illustrates why robustness and verifiability are not optional performance enhancements but central safety requirements (Zhou et al., 2024). Clinical environments include complex language artifacts such as progress notes, discharge summaries, radiology reports, medication lists, and patient communications, which are often incomplete and contain contradictions. LLM systems can support clinicians by summarizing histories, drafting patient-friendly explanations, extracting structured variables, assisting with documentation, and surfacing guideline-relevant considerations. However, the same language fluency that enables these functions can produce confident-sounding statements that are subtly wrong, clinically unsafe, or poorly grounded in a patient's record (Mubarak et al., 2023). Robustness in healthcare requires that a system behaves consistently across note styles, institutions, and patient subgroups; that it resists misleading inputs; and that it remains reliable when information is missing or uncertain. Verifiability in healthcare requires that any recommendation or claim can be traced to evidence in the chart or authoritative references, that uncertainty is expressed in a way clinicians can interpret, and that the system supports abstention when confidence is low. Clinical decision-making also highlights human factors: clinicians may over-trust outputs under time pressure, and documentation generated by an LLM may become part of the medical record, reinforcing errors through later reuse (Revell et al., 2024). Therefore, robust and verifiable healthcare LLMs must be designed as systems that prioritize traceability, calibrated uncertainty, and safe escalation pathways to human review. These requirements are amplified internationally because clinical guidelines, formularies, and languages vary across regions, and safety assurance must remain meaningful under cross-jurisdiction differences in practice and regulation.

Defense and national security contexts place additional emphasis on adversarial robustness, operational security, and controlled information flows. LLMs can be used to synthesize intelligence reporting, assist with planning and logistics, summarize communications, support threat analysis, and accelerate routine administrative tasks (Aharoni et al., 2024; Muhammad Mohiul, 2020). The risk profile is distinct because adversaries can intentionally manipulate inputs, exploit vulnerabilities in tool-using systems, and attempt to induce policy violations, misinformation, or disclosure of sensitive information (Jinnat & Md. Kamrul, 2021). Robustness in defense settings therefore includes resistance to prompt injection, deception, and crafted documents designed to hijack model behavior. It also includes stability under rapidly changing contexts, incomplete evidence, and high ambiguity, where the system must avoid prematurely converging on a single narrative (Hasan & Shaikat, 2021; Mirzaei et al., 2024). Verifiability in defense settings requires auditable decision logs, clear provenance for claims, access control enforcement, and reproducible evaluation protocols that demonstrate behavior under adversarial stress (Rabiul & Samia, 2021). A decision-support LLM must be able to show what evidence it relied upon, which constraints were applied, and where uncertainty remained. Because defense operations often involve coalitions and international partnerships, verifiability also needs to support interoperable governance: shared standards for audit trails, testing, and secure deployment across multiple organizations (Mohiul & Rahman, 2021; Zhang et al., 2020). This international dimension increases the importance of consistent measurement and assurance practices, since trust between partners depends on the ability to validate system behavior and ensure that information-sharing agreements and security policies are respected (Rahman & Abdul, 2021).

Figure 1: Robust Verifiable High-Stakes LLM Engineering



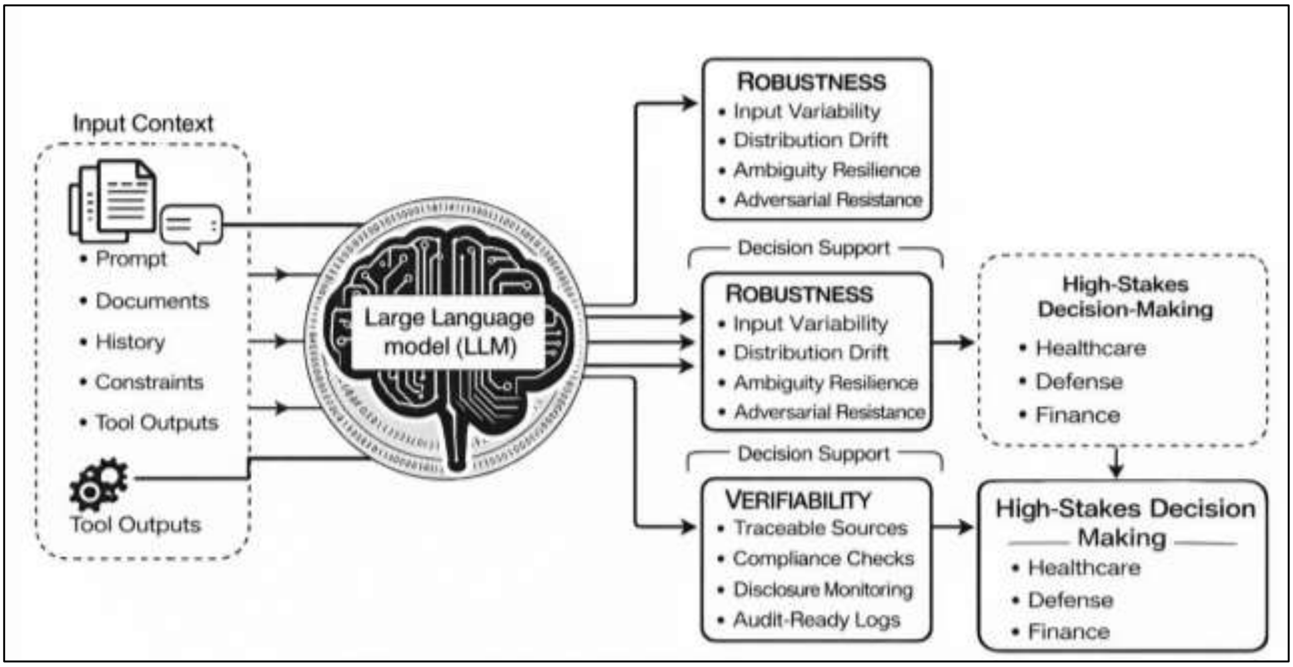
Finance similarly demands robustness and verifiability because decisions are tightly coupled with regulation, auditability, and systemic risk (Luo et al., 2024; Haider & Shahrin, 2021). Financial institutions use language and documents as core operating artifacts: customer communications, disclosures, analyst reports, policy documents, compliance rules, transaction narratives, and market news (Zulqarnain & Subrato, 2021). LLMs can assist with customer service, fraud analysis, contract review, compliance screening, summarization for analysts, and generation of internal reports. The harm from errors can be immediate and large: misstatements can trigger regulatory violations, erroneous advice can lead to customer losses, flawed compliance decisions can incur fines, and mistakes in risk assessment can amplify instability. Robustness in finance includes stable performance under market regime shifts, changes in product offerings, evolving regulatory requirements, and adversarial attempts to bypass controls (Goeuriot et al., 2024; Habibullah & Farabe, 2022; Arman & Kamrul, 2022). Verifiability includes traceable sources for any factual statements, explicit alignment with compliance rules, and documentation suitable for internal model governance processes. A verifiable system should support controlled abstention, flagging uncertain cases for human review, and maintaining logs that allow auditors to reconstruct what the system did and why (Rashid & Sai Praveen, 2022; Kamrul & Omar, 2022). International significance is pronounced because financial systems are interconnected across borders, regulations differ by region, and many institutions operate in multilingual environments. A robust and verifiable LLM that can be audited across jurisdictions and can demonstrate consistent governance behavior becomes a prerequisite for responsible adoption at scale (Kumar & Chand, 2020; Rahman, 2022; Rony & Samia, 2022).

A central technical challenge to robust deployment is the mismatch between fluent generation and reliable truth. LLM outputs are optimized to be plausible continuations, not guaranteed factual statements (Abdul & Rahman, 2023; Aditya & Rony, 2023; Kumar & Chand, 2021). This creates multiple observable failure modes in high-stakes settings: hallucinations (fabricated facts), brittle sensitivity to prompt phrasing, inconsistent answers across paraphrases, and susceptibility to adversarial instructions embedded in inputs or retrieved documents. Robustness requires quantitative characterization of these failures under realistic perturbations: changes in wording, insertion of distractor content, missing information, contradictory evidence, and adversarial prompts (Arfan & Rony, 2023; Ara & Shaikh, 2023). It also requires measuring the stability of outputs when the same case is presented with alternative phrasing or different ordering of evidence. Another core requirement is calibrated uncertainty: the system should not only produce an answer but also produce a meaningful estimate of confidence or risk, enabling selective prediction and abstention (Freitas et al., 2020; Habibullah & Mohiul, 2023; Hasan & Waladur, 2023). In high-stakes domains, abstention is a safety mechanism: refusing to answer or escalating to human review when uncertainty is high can reduce harmful errors, even if it reduces coverage. Robustness also involves security-aware design. Tool-using LLMs that retrieve documents, call external services, or execute code expand the attack surface; malicious content can attempt to alter instructions, leak secrets, or induce unsafe actions (Arman & Nahid, 2023; Mesbaul, 2023). Therefore, robust high-stakes systems must incorporate secure retrieval, content sanitization, strict separation between untrusted inputs and system instructions, and evaluation protocols that include adversarial testing. These technical requirements translate naturally into quantitative variables for a paper: error rates under perturbations, hallucination incidence, inconsistency metrics across paraphrases, uncertainty calibration scores, abstention performance curves, and adversarial success rates (Jia et al., 2022; Milon & Mominul, 2023; Mohaiminul & Muzahidul, 2023).

Verifiability requires turning an LLM's behavior into something measurable, auditable, and reproducible (Musfiqur & Kamrul, 2023; Rezaul & Kamrul, 2023; Sangeetha et al., 2024). One practical route is evidence grounding: constraining outputs so that key claims are linked to identifiable sources, such as documents in a patient record, policy manuals, financial filings, or approved intelligence references (Amin & Praveen, 2023; Rabiul & Mushfequr, 2023). When the system is designed so that answers must be supported by retrieved evidence, auditing becomes feasible: reviewers can check whether the cited evidence actually supports the claim, whether the evidence is current, and whether critical counter-evidence was ignored. A second route is modular tool use, where computations and rule checks are delegated to deterministic components. For example, a finance assistant might use a rules engine for compliance constraints, while an LLM provides natural language explanation constrained by the rules output (Shahrin & Samia, 2023; Roy, 2023; Qayyum et al., 2020). A third route is formal verification for components of the pipeline: while verifying an entire generative system end-to-end is difficult, it is more realistic to verify properties of the surrounding system such as access control logic, transaction constraints, or critical rule enforcement (Rakibul & Majumder, 2023; Rifat & Rebeka, 2023). Verifiability also includes reproducibility: the ability to recreate system outputs given the same inputs and configuration, which is important for audits and incident investigations (Sabuj Kumar, 2023; Saikat & Aditya, 2023). This pushes design toward controlled decoding settings, versioned models, immutable logs, and deterministic tool outputs where possible (Rashid, 2024; Zulqarnain & Subrato, 2023). For a quantitative study, verifiability can be operationalized with measurable indicators: citation precision and recall, evidence coverage, rate of unsupported claims, reproducibility under fixed seeds and configurations, compliance-check pass rates, and audit-trail completeness (Ansari et al., 2020; Md & Praveen, 2024; Mohaiminul & Majumder, 2024). These measures can be analyzed across domains to compare how constraints differ and which architectures provide stronger assurance.

Rigorous evaluation is the method that ties robustness and verifiability to measurable claims appropriate for high-stakes adoption. Traditional benchmark accuracy is insufficient because it does not measure behavior under distribution shift, adversarial attempts, or operational constraints (Javed et al., 2024; Foysal & Abdulla, 2024; Ibne & Aditya, 2024). A quantitative evaluation framework for robust and verifiable LLMs should include stress testing across input perturbations, multilingual and cross-institution variations, and domain-specific threat models (Milon & Mominul, 2024; Mosheur & Arman, 2024). It should include scenario-based testing where an LLM must integrate evidence from multiple sources, handle contradictions, and decide whether to answer or abstain. It should also incorporate human-in-the-loop evaluation because the real-world impact of an LLM depends on how professionals interpret outputs, how often they detect errors, and whether the system increases or decreases overall decision quality (Rahman & Aditya, 2024; Saba & Hasan, 2024). In healthcare, that may involve measuring whether clinicians identify unsafe recommendations; in finance, whether compliance officers can audit outputs efficiently; in defense, whether analysts detect deception attempts embedded in inputs (Natarajan et al., 2023; Kumar, 2024; Praveen, 2024).

Figure 2: Robust Verifiable LLMs for Decisions



Evaluation must also measure tradeoffs: stricter guardrails can reduce unsafe outputs while increasing unnecessary refusals; retrieval grounding can reduce hallucinations while increasing dependence on retrieval quality; abstention can reduce harm while reducing coverage. Because high-stakes environments are internationally distributed, evaluation should also consider cross-jurisdiction data, language variants, and policy differences (Praveen, 2024; Shaikat & Aditya, 2024). Quantitative comparisons can be structured using consistent metrics across domains—robustness under shift, calibration and abstention quality, evidence grounding fidelity, adversarial resilience, and auditability—while still respecting domain-specific constraints (Arfan, 2025; Ara, 2025). This framing motivates a measurement-driven approach where robust and verifiable LLMs are assessed by multi-dimensional reliability profiles rather than by single performance numbers (Erdemir et al., 2020; Jinnat, 2025; Rashid, 2025b).

The objective of the quantitative study is to empirically measure and compare the reliability, robustness, and verifiability of large language model-based decision-support systems under conditions that realistically represent high-stakes operational environments across three critical domains . Specifically, the study aims to quantify how consistently LLM outputs remain correct and stable when inputs are perturbed through clinically and operationally plausible variations such as paraphrasing, incomplete records, contradictory evidence, domain-specific jargon, and adversarially crafted instructions, while also measuring the system’s ability to appropriately abstain or escalate cases when confidence is insufficient. A central objective is to develop and apply a multi-metric evaluation framework that simultaneously captures (a) decision accuracy and task utility, (b) robustness to distribution shift and input noise, (c) calibration of uncertainty and risk-coverage tradeoffs under selective prediction, (d) evidence-grounding quality through traceable support for claims, and (e) auditability through reproducibility and log completeness. The study further aims to compare different system configurations—such as baseline prompting, retrieval-grounded generation, rule- or tool-augmented pipelines, and constraint-enforced response formats—to determine which design choices produce statistically significant improvements in robustness and verifiability without reducing operational usability. Another objective is to identify domain-sensitive failure patterns by analyzing error types and instability modes separately in healthcare (e.g., clinical hallucinations and unsafe recommendations), defense (e.g., susceptibility to manipulation and policy violations), and finance (e.g., compliance inconsistencies and factual inaccuracies in regulated content), enabling cross-domain comparison of risk profiles using a common measurement protocol. Finally, the study aims to generate quantitative evidence that can support standardized assurance reporting by producing measurable thresholds, confidence intervals, and comparative performance rankings for robustness and verifiability indicators, thereby enabling organizations to evaluate whether an LLM-based decision-support system meets minimum reliability requirements for high-stakes deployment across international and multi-jurisdiction contexts.

LITERATURE REVIEW

The literature review for "Robust and Verifiable LLMs for High-Stakes Decision-Making (Healthcare, Defense, Finance)" synthesizes empirical and methodological research that explains how large language model (LLM) systems can be evaluated and engineered to meet measurable assurance requirements in environments where decision errors create severe harm, regulatory exposure, or systemic instability (Harvey, 2024). In high-stakes domains, conventional performance reporting (single benchmark accuracy or generic helpfulness scores) is insufficient because real-world deployment involves noisy and incomplete inputs, institutional variation, distribution shifts, adversarial manipulation, and strict audit expectations. Accordingly, this literature review is organized around two measurable constructs: robustness, defined as stable and safe performance under perturbations, shifts, and adversarial conditions; and verifiability, defined as the extent to which outputs and system behaviors can be checked through traceable evidence, constraint compliance, reproducibility, and audit-ready logs (Felderer & Ramler, 2021). The review prioritizes studies that operationalize these constructs quantitatively through reliability metrics, calibration measures, risk-coverage analyses, adversarial success rates, evidence-grounding fidelity, and reproducibility indicators. It further integrates domain-specific evidence from healthcare, defense, and finance to highlight how assurance requirements differ by consequence structure, threat model, and governance constraints, while still supporting a unified evaluation protocol. The section closes by consolidating prior findings into a measurement framework that directly informs the present study's variables, hypotheses, and statistical analysis plan, enabling cross-domain comparison of LLM configurations (baseline prompting, retrieval-grounded generation, tool-augmented pipelines, and rule-constrained outputs) using standardized quantitative reporting (Sim et al., 2021).

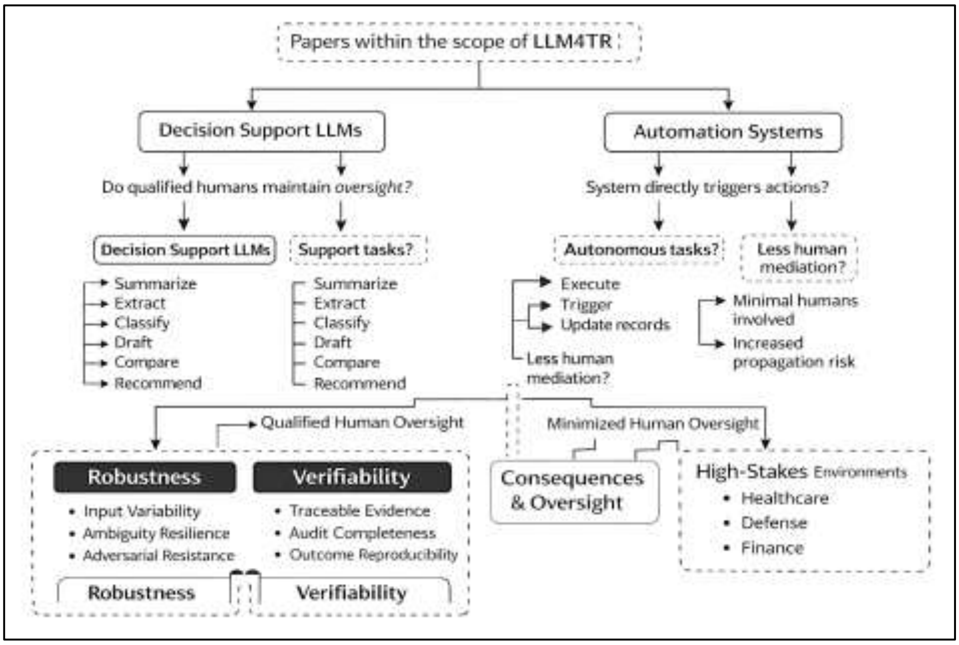
Assurance Problem and Core Constructs

Large language models (LLMs) have increasingly been positioned as *decision-support* technologies rather than autonomous decision makers, and the literature draws a practical boundary between these roles by focusing on where responsibility and control reside in the workflow (Hildesheim & Sonntag, 2020; Md Harun-Or-Rashid, 2025a; Md Mesbaul, 2025). In decision support, the LLM is used to summarize, extract, classify, draft, compare, and recommend while a qualified human actor retains authority to accept, reject, or revise the output and remains accountable for the decision record. In automation, the system executes or triggers actions with minimal human mediation, which raises the assurance burden because failures can propagate faster and with fewer opportunities for interception. Research across applied clinical NLP, operational risk management, and safety-critical human–automation systems emphasizes that the difference is not semantic (Md. Milon, 2025; Md. Mosheur, 2025); it changes how errors are detected, how audit trails are constructed, and how liability and oversight are assigned. In high-stakes contexts, the literature consistently frames "stakes" through consequence magnitude and oversight burden. Consequence magnitude refers to the severity and irreversibility of harm that can follow from errors—patient injury, financial loss, regulatory breach, security compromise, or cascading institutional disruption (Rabiul, 2025; Shahrin, 2025; Taylor et al., 2020). Oversight burden refers to the degree of documentation, validation, traceability, and governance expected before and after a system is used, including ongoing monitoring and incident response. Together, these dimensions motivate an assurance problem: LLMs produce fluent language that can look authoritative while still being incorrect, incomplete, or poorly grounded. Prior studies of factuality, hallucination, and instability show that language plausibility is not a reliable proxy for truth or safety, and high-stakes deployments intensify this gap because users operate under time pressure and may not independently verify each claim. The literature also highlights that assurance for LLMs must be defined at the system level rather than the model level, because "LLM output" is shaped by prompt protocols, retrieval components, tool integrations, decoding settings, and user interaction patterns. Accordingly, the assurance problem is commonly defined as the challenge of ensuring that a decision-support LLM remains reliable under realistic operating conditions, communicates uncertainty appropriately, and produces outputs that can be checked and reconstructed through evidence and logs (Almeida et al., 2022). This framing sets the stage for operational constructs that are repeatedly emphasized across studies: robustness, referring to stable and safe behavior under stressors, and verifiability, referring to the degree to which outputs and system actions can be checked, audited, and reproduced.

Within this assurance framing, the literature treats robustness as a property that must be defined operationally in relation to the conditions that cause performance to degrade in the real world (Channuntapipat et al., 2020). Robustness is discussed as stability under input variability such as paraphrasing, formatting changes, incomplete or noisy records, contradictory information, domain-specific jargon, and shifts in how institutions document cases. It is also discussed as resilience to adversarial manipulation, especially when the LLM is embedded in workflows that ingest untrusted text such as emails, web pages, or uploaded documents. Studies that examine brittleness in question answering, summarization, and professional-domain tasks show that small changes in instruction wording can change answers, alter rationales, or shift confidence tone, creating inconsistency that is operationally costly in audited environments. The literature further connects robustness to reliability under distribution shift, because healthcare, defense, and finance each contain substantial cross-site variation: hospitals differ in clinical note style and coding, defense reporting differs by unit and classification context, and financial institutions differ by product line, jurisdiction, and compliance rules. Robustness is therefore treated as multi-dimensional: it includes correctness stability, consistency of outputs across equivalent inputs, resistance to manipulation, and graceful handling of missing evidence. In quantitative assurance discussions, robustness is represented through measurable indicators such as performance degradation under controlled perturbations, inconsistency rates across paraphrases, stability across different sources, and the frequency of unsafe or policy-incompatible behaviors under stress tests (Rakibul, 2025; Kumar, 2025; Sumi & Kabir, 2021). This literature also places strong emphasis on the shape of failures rather than only the average error rate. Rare but severe errors, particularly those expressed with confident language, are repeatedly identified as disproportionate drivers of risk. As a result, robustness is commonly linked with calibrated uncertainty behavior, where a robust system signals uncertainty or abstains in cases where evidence is insufficient. However, the core definitional point that unifies prior work is that robustness is not simply "high accuracy"; it is "high reliability when the input distribution and interaction conditions change," including conditions that are normal in operations but underrepresented in benchmarks. This definition is essential for high-stakes adoption because it anchors evaluation to the kinds of variability that professional users and regulators expect a system to survive without producing brittle or misleading outputs (Sai Praveen & Md, 2025; Studer et al., 2021).

The literature defines verifiability as the ability to subject an LLM system's outputs and behaviors to checking procedures that are independent of the model's own narrative. Verifiability is presented as an antidote to the risk that persuasive language can substitute for evidence. In professional domains, checking can take multiple forms: linking claims to source documents, ensuring that recommendations map to explicit rules or guidelines, validating numeric statements via deterministic tools, and preserving logs that allow reviewers to reproduce the context and reconstruct what the system saw and did (Hobbs et al., 2023). Researchers repeatedly distinguish verifiability from generic explainability. Explanations can be helpful for users, but prior studies show that explanations can also be fabricated or post-hoc rationalizations that increase trust without increasing correctness. Verifiability, in contrast, requires that claims be grounded in inspectable artifacts such as cited excerpts, structured evidence tables, tool outputs, or rule-check results.

Figure 3: Decision Support vs Automation LLMs



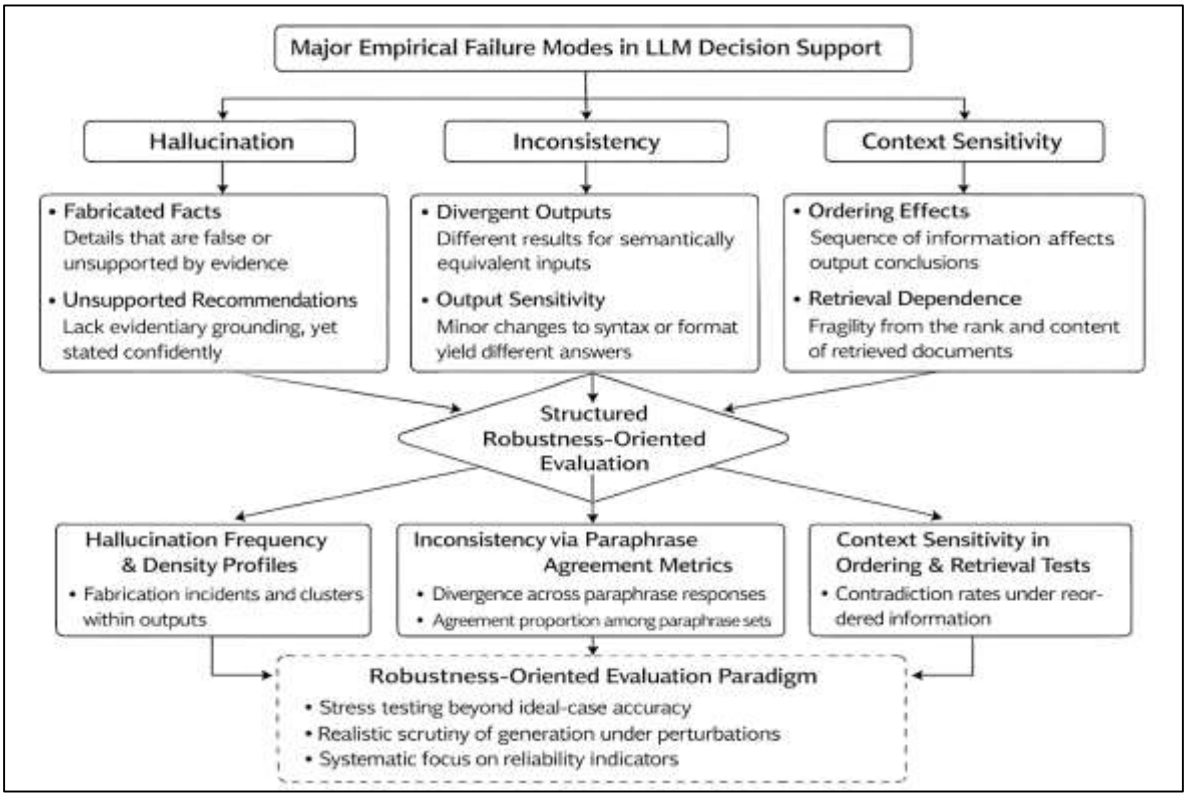
In system design research, retrieval-grounded generation is often discussed as a practical route to verifiability because it can attach outputs to identifiable sources, enabling reviewers to confirm whether the evidence actually supports the claim. Rule-constrained output formats and schema validation are also discussed as verifiability mechanisms because they reduce ambiguity and make omissions detectable, which is important when documentation must meet legal or clinical standards (Almuhaideb & Saeed, 2020). The literature also emphasizes reproducibility as a verifiability dimension: audited environments require versioning of models, prompts, retrieval indexes, and policies; deterministic or controlled generation settings; and complete record-keeping of inputs, retrieved material, and tool calls. Quantitative discussions of verifiability therefore rely on measurable indicators such as the proportion of claims supported by evidence, the correctness of source attributions, the completeness of audit logs, the consistency of outputs under reruns with fixed configurations, and the rate at which outputs meet required structural constraints. In high-stakes settings, this is closely tied to governance practices that demand technical documentation, risk controls, and incident reporting procedures, which collectively make “verifiability” a system property spanning both engineering and process (Bylund & Packard, 2022). The literature’s consistent conclusion at the definitional level is that verifiability is achieved when a competent reviewer can check the output against evidence, understand which constraints governed the response, and reproduce the decision-support artifact from recorded inputs and configurations.

A final body of literature synthesizes robustness and verifiability into measurable boundaries suitable for comparing performance across healthcare, defense, and finance, while acknowledging that the meaning of “acceptable” varies by domain (Bile Hassan et al., 2022). Domain mapping is typically accomplished by introducing a risk-tier scheme that groups tasks according to consequence severity, required oversight, and adversarial exposure. In healthcare, tasks involving medication guidance, triage, diagnosis support, or patient harm are commonly treated as higher risk than tasks involving administrative drafting, coding suggestions, or generic education summaries. In finance, tasks that affect credit decisions, compliance determinations, customer advice, or market-sensitive reporting are treated as higher risk than tasks limited to internal summarization or non-actionable insights. In defense, tasks related to threat assessment, operational planning, secure communications, and intelligence synthesis carry higher risk than tasks related to routine administrative support. Across these domains, the literature suggests that risk tiering is necessary because it aligns evaluation intensity with consequence magnitude and prevents misleading comparisons between low-impact and high-impact use cases (Abbas, 2020). This mapping also clarifies which robustness threats dominate: healthcare emphasizes variability in documentation, missingness, and clinical safety; finance emphasizes regulatory constraints, auditability, and regime shifts; defense emphasizes adversarial manipulation, secrecy, and strict policy enforcement. In measurable terms, robustness indicators are interpreted differently across tiers: a small error rate may still be unacceptable for the highest tiers, and inconsistency may be more damaging when outputs must be repeatable for audit. Verifiability indicators likewise shift: evidence grounding and traceability may be essential for top-tier tasks across all domains, while reproducibility and complete logging become non-negotiable where external audits and incident reviews are expected. The literature therefore supports an assurance framework that defines robustness and verifiability not as abstract ideals but as operational constructs linked to task risk tiers and domain constraints (Humphrey et al., 2021). This creates a structured boundary for the literature review: LLM decision support must be evaluated by how it behaves under realistic stressors and by how thoroughly its outputs can be checked and reconstructed, with risk tiering providing the organizing principle for selecting metrics, test conditions, and acceptable thresholds in healthcare, defense, and finance.

Empirical Failure Modes That Motivate Robustness

Large language model decision-support systems exhibit a persistent class of empirical failures that undermine their reliability in high-stakes environments, most notably through hallucination behaviors that manifest as fluent but incorrect content. Hallucinations appear in multiple operationally relevant forms (Zhang & Ng, 2021). Fabricated facts occur when a model introduces details that are absent from the input context and unsupported by any underlying evidence, often filling informational gaps with plausible but false statements. In professional decision-support workflows, these fabrications are especially problematic because they can blend seamlessly with legitimate content, making detection difficult without systematic verification. Unsupported recommendations represent a second hallucination type, in which a system moves beyond summarization or analysis into prescriptive guidance without sufficient evidentiary grounding. This is particularly hazardous in healthcare, defense, and finance, where recommendations may influence treatment plans, operational judgments, or compliance actions. A third category involves incorrect numerical information, including wrong quantities, thresholds, percentages, dates, or monetary figures (Zhong & Liu, 2024). Such numerical hallucinations are frequently expressed with unwarranted precision, creating a false sense of accuracy that can mislead users into trusting incorrect outputs. Across empirical evaluations, hallucinations are shown to occur not only as isolated mistakes but as clusters within a single response, where multiple unsupported statements reinforce one another. This compounding effect increases the likelihood that an output will be accepted as valid, especially under time pressure or cognitive load. Consequently, empirical research treats hallucination as a measurable reliability signal rather than an anecdotal error, examining how often hallucinations occur within decision cases and how densely they appear within generated text. This framing reflects a broader consensus that hallucination frequency and severity are central indicators of system robustness, as they reveal the degree to which a model prioritizes linguistic plausibility over factual grounding. In high-stakes settings, the cost of hallucinations is amplified because outputs may be incorporated into records, reports, or downstream analyses, transforming a single model error into a persistent organizational artifact (Han et al., 2024). These findings motivate robustness as a core requirement, defined not only by correctness in ideal conditions but by resistance to producing confident, unsupported content when evidence is incomplete, ambiguous, or conflicting. Beyond hallucination, inconsistency across paraphrases and prompt templates emerges as a second dominant empirical failure mode motivating robustness requirements. Inconsistency occurs when semantically equivalent inputs yield different outputs, conclusions, or recommendations, even though the underlying task remains unchanged (Steenhoek et al., 2023). Empirical studies repeatedly demonstrate that minor variations in wording, formatting, or instruction ordering can lead to answer reversals, altered reasoning paths, or shifts in emphasis. This sensitivity is particularly problematic in high-stakes decision support because professional environments demand repeatability and defensibility. A decision-support system that produces different answers for the same case under slightly different phrasing undermines auditability and weakens institutional trust. Inconsistency also complicates governance processes, as reviewers cannot easily determine whether divergent outputs reflect meaningful uncertainty or arbitrary sensitivity to prompt structure. The literature shows that inconsistency is not limited to final answers; it also affects intermediate elements such as selected evidence, extracted entities, prioritization of risks, and confidence tone (Finlay & Oberman, 2021). As a result, inconsistency is increasingly treated as a structured failure mode rather than random noise. Measurement approaches focus on the proportion of cases that exhibit divergent outputs across repeated trials and the degree of agreement among outputs generated from paraphrase sets. Importantly, inconsistency is shown to persist even in models optimized for instruction following and conversational alignment, indicating that training improvements alone do not eliminate instability. In high-stakes domains, this instability interacts with human judgment in complex ways. Users may selectively trust outputs that align with expectations, disregard conflicting results, or lose confidence in the system altogether. Each of these outcomes can degrade decision quality. Therefore, inconsistency is not merely a technical concern but a socio-technical risk factor that affects how LLM outputs are interpreted and used (Weber et al., 2023). The literature's emphasis on consistency as a measurable property reflects the need to evaluate robustness at the level of repeated, equivalent decision-support interactions rather than single-run performance, reinforcing the importance of systematic robustness testing under controlled variation.

Figure 4: Empirical Failure Modes in LLMs


Major Empirical Failure Modes in LLM Decision Support
Hallucination
Inconsistency
Context Sensitivity
• Fabricated Facts
Details that are false or unsupported by evidence
• Unsupported Recommendations
Lack evidentiary grounding, yet stated confidently
• Divergent Outputs
Different results for semantically equivalent inputs
• Output Sensitivity
Minor changes to syntax or format yield different answers
• Ordering Effects
Sequence of information affects output conclusions
• Retrieval Dependence
Fragility from the rank and content of retrieved documents
Structured Robustness-Oriented Evaluation
Hallucination Frequency & Density Profiles
• Fabrication incidents and clusters within outputs
Inconsistency via Paraphrase Agreement Metrics
• Divergence across paraphrase responses
• Agreement proportion among paraphrase sets
Context Sensitivity in Ordering & Retrieval Tests
• Contradiction rates under reordered information
Robustness-Oriented Evaluation Paradigm
• Stress testing beyond ideal-case accuracy
• Realistic scrutiny of generation under perturbations
• Systematic focus on reliability indicators

A third cluster of empirical failures arises from context sensitivity, which includes ordering effects, retrieval dependence, and long-context drift. Ordering effects occur when the sequence in which information or constraints is presented influences the system's conclusions, even when the content remains constant. In decision-support scenarios involving multiple documents or complex instructions, the same facts can yield different outputs depending on their position in the prompt or the relative prominence of constraints. This behavior reflects limitations in how LLMs integrate and prioritize information across extended contexts (Freiesleben & Grote, 2023). Retrieval dependence introduces an additional layer of fragility in systems that rely on external documents. While retrieval can improve grounding, empirical evaluations show that outputs become highly contingent on which documents are retrieved, how they are ranked, and whether conflicting or outdated sources are included. When retrieval surfaces irrelevant or misleading material, the model may anchor on that content and generate outputs that appear grounded but are substantively flawed. This creates a new class of failure in which errors originate not from the model alone but from interactions between retrieval quality and generation behavior. Long-context drift further complicates reliability as context length increases (Ferdous et al., 2020). As more information is introduced, models may omit critical constraints, conflate entities, or lose track of earlier instructions, resulting in subtle contradictions or incomplete reasoning. These errors are difficult to detect because the output may remain coherent and professionally styled. Empirical research emphasizes that context sensitivity is especially relevant in healthcare, finance, and defense, where decision cases often involve long histories, multiple sources, and evolving evidence. Measurement approaches therefore focus on how outputs change when evidence order is altered, when retrieval sets differ, or when context length expands (Heinze-Deml & Meinshausen, 2021). Contradiction rates under conflicting evidence and omission patterns under long contexts are treated as key indicators of robustness. This body of literature reinforces the view that robust decision-support systems must demonstrate stability across different representations of the same information, rather than relying on a single canonical prompt structure.

Taken together, these empirical failure modes motivate a robustness-oriented evaluation paradigm that treats hallucination, inconsistency, and context sensitivity as interconnected and measurable dimensions of risk (Liu et al., 2021). The literature consistently argues that evaluating LLM decision support requires more than assessing correctness under ideal conditions. Instead, reliability must be characterized across repeated trials, varied inputs, and realistic operational stressors. Hallucinations are examined in terms of frequency, density, and actionability within outputs, recognizing that not all errors carry equal risk. Inconsistency is evaluated through agreement metrics across paraphrases and prompt variants, capturing stability rather than average performance. Context sensitivity is assessed by systematically varying evidence order, retrieval inputs, and context length to identify conditions under which outputs diverge or contradict available information (Mao et al., 2020). These measurement practices reflect a shift from benchmark-centric evaluation toward reliability profiling, where the goal is to map how and when systems fail. This shift is particularly salient in high-stakes domains because rare but severe failures can dominate risk profiles. The literature also highlights that these failure modes interact with human use patterns, as persuasive language and professional formatting can mask underlying instability or lack of evidence. As a result, robustness is framed not simply as a technical attribute of models but as a system-level property that determines whether decision-support outputs remain trustworthy under the pressures of real-world use. By consolidating these empirical findings, the literature establishes a clear motivation for treating robustness as a core assurance requirement and for grounding its assessment in structured, repeatable measurements that capture how decision-support systems behave when conditions deviate from the ideal (Abdollahi & Ebrahimi, 2020).

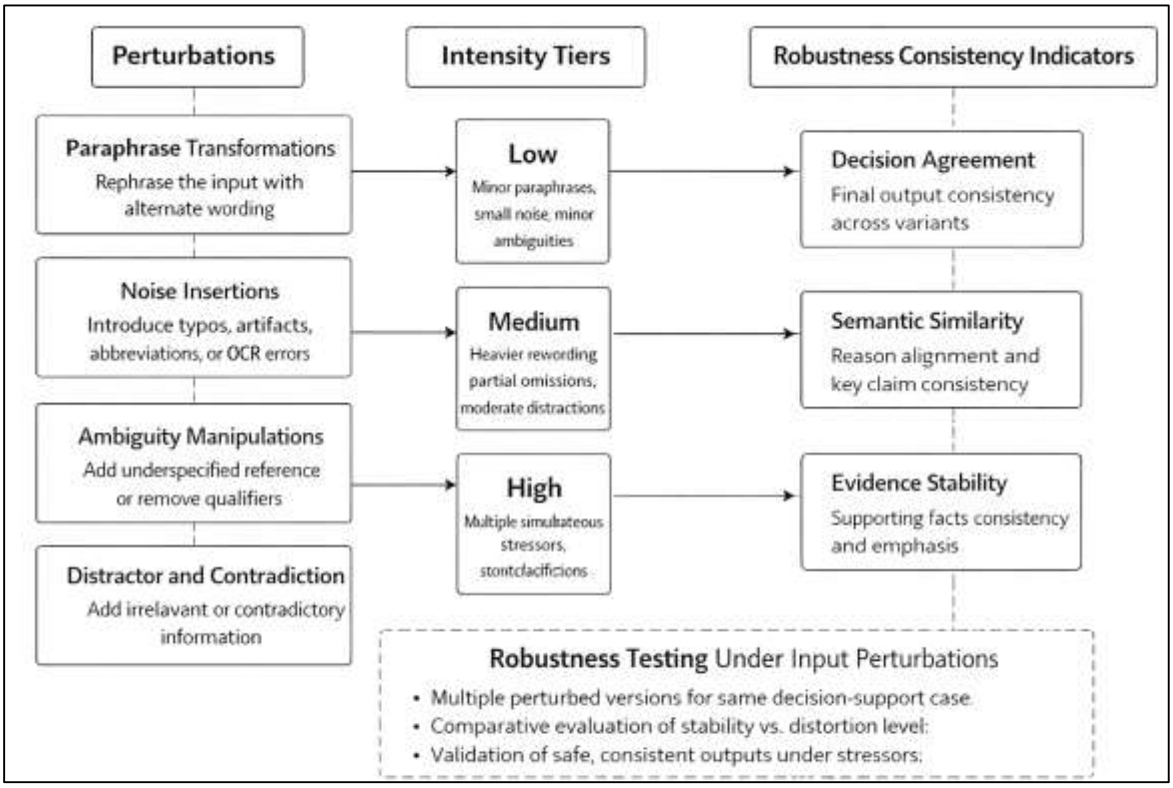
Robustness Testing Under Input Perturbations

Robustness testing under input perturbations is widely treated in the literature as a necessary complement to standard benchmark evaluation because high-stakes decision support operates in environments where inputs are rarely clean, uniform, or perfectly specified (Mahmoud et al., 2020). Research across adversarial NLP, reliability testing, and domain-specific evaluation shows that a model's apparent competence can mask sensitivity to superficial changes in wording, formatting, and contextual packaging. Consequently, perturbation-based testing is used to quantify how stable the same underlying decision-support task remains when the input is altered in ways that preserve the intended meaning or reflect realistic documentation variability. Perturbation sets commonly include paraphrase transformations that restate a request or evidence in alternative wording, noise insertion such as typos, abbreviations, OCR-like artifacts, or inconsistent punctuation, and ambiguity manipulations that introduce underspecified references or missing qualifiers. The literature also emphasizes distractor perturbations that add irrelevant but plausible content and contradiction perturbations that insert conflicting evidence to examine whether a model can reconcile inconsistencies or erroneously commit to a single narrative (Mechali et al., 2021). These perturbations are treated as essential stressors because they map to actual conditions in healthcare notes, finance filings, customer communications, and intelligence reporting, where documents include incomplete segments, mixed terminology, and contested claims. Robustness testing frameworks therefore adopt a repeated-measures philosophy: the same case is presented across multiple perturbed variants so that stability is evaluated as a property of the system, not as a single-run outcome. This perspective aligns with broader evaluation research suggesting that robust systems should preserve core decisions under semantically equivalent input changes, demonstrate controlled behavior under ambiguous prompts, and maintain safety constraints when distractors or contradictions are present. As a literature-driven approach, perturbation testing is presented as both a diagnostic tool for understanding failure patterns and a quantitative method for comparing system configurations such as baseline prompting, retrieval-grounded pipelines, or constrained output formats. In high-stakes contexts, this approach is motivated by the observation that decision-support errors often arise from ordinary variability in records rather than from rare corner cases, and perturbation testing provides a structured way to simulate that variability while preserving experimental control (Rusak et al., 2020).

A central methodological theme in perturbation-based robustness research is the use of controlled intensity tiers that represent increasing degrees of input distortion, allowing evaluation to capture not only whether a system fails but how rapidly its reliability deteriorates as conditions become harder (Nielsen et al., 2022). The literature commonly operationalizes intensity in graded forms: low-intensity perturbations may involve minor paraphrasing, small typos, or slight reformatting; medium-intensity perturbations may include heavier rewording, added distractor sentences, partial omissions, or moderate ambiguity; high-intensity perturbations may include multiple simultaneous stressors such as contradictory evidence combined with noise, deeply ambiguous queries, or substantial distractor insertion that mimics real-world clutter. This tiering approach is valuable in high-stakes domains because it provides a more interpretable picture of stability than binary pass/fail results. A system that performs well on clean inputs but collapses under moderate perturbations is empirically different from a system that maintains stable outputs until high-intensity distortions (Labbadi & Cherkaoui, 2021). Studies of robustness commonly compare performance curves across tiers to characterize resilience patterns and to locate thresholds where reliability becomes unacceptable for a given risk tier. This line of work also emphasizes that intensity design must remain domain-faithful: in healthcare, intensity should reflect real chart noise, note fragmentation, and abbreviation density; in finance, it should reflect filing verbosity, mixed numeric reporting, and policy-rule language; in defense, it should reflect compressed reporting, code words, and conflicting situational updates. The literature also notes that intensity tiers allow researchers to separate brittle sensitivity from meaningful uncertainty. If outputs shift dramatically under low-intensity paraphrases, that suggests instability rather than evidence-driven revision. If shifts occur primarily under high-intensity contradictions, that may reflect the system's difficulty resolving genuine evidence conflict. Robustness research therefore uses tiered perturbations to produce comparative stability profiles and to examine whether improvements in one dimension, such as grounding via retrieval, reduce degradation under certain stressors while leaving others unchanged. This tiered structure supports quantitative study designs that compare multiple systems or configurations under identical perturbation schedules, yielding reliable comparisons across domains and task types (Meng et al., 2022).

Another major focus in the literature is the measurement of stability under perturbations using consistency indicators that capture not only accuracy but also agreement, semantic invariance, and contradiction handling (Gandhi & Jain, 2020). Robustness is assessed through measures that compare clean-input outputs to perturbed-input outputs for the same case, evaluating whether the decision outcome remains aligned with the original and whether the reasoning content remains consistent. Consistency is treated as multi-level: decision agreement captures whether the final recommendation or classification remains the same; semantic similarity captures whether the explanation content and key claims remain aligned; and evidence stability captures whether the same supporting facts are selected and emphasized across variants. This approach reflects a broader research consensus that high-stakes decision support cannot rely only on final-answer correctness, because the rationale and supporting content often influences user trust, audit interpretation, and downstream actions. Research on contradiction perturbations further emphasizes the need to evaluate whether systems recognize evidence conflict, avoid forced certainty, and appropriately qualify conclusions when presented with inconsistent inputs (Chongzhi Zhang et al., 2020).

Figure 5: Robustness Testing Under Input Perturbations


Perturbations
Intensity Tiers
Robustness Consistency Indicators
Paraphrase Transformations
Rephrase the input with alternate wording
Low
Minor paraphrases, small noise, minor ambiguities
Decision Agreement
Final output consistency across variants
Noise Insertions
Introduce typos, artifacts, abbreviations, or OCR errors
Medium
Heavier rewording partial omissions, moderate distractions
Semantic Similarity
Reason alignment and key claim consistency
Ambiguity Manipulations
Add underspecified reference or remove qualifiers
High
Multiple simulbateous stressors, stontclacfiicions
Evidence Stability
Supporting facts consistency and emphasis
Distractor and Contradiction
Add irrelavant or contradictory information
Robustness Testing Under Input Perturbations
• Multiple perturbed versions for same decision-support case
• Comparative evaluation of stability vs. distortion level:
• Validation of safe, consistent outputs under stressors:

In such scenarios, stability does not necessarily mean producing identical outputs; it means demonstrating controlled behavior such as consistently identifying the contradiction, maintaining safety constraints, and avoiding unsupported resolution. The literature also highlights that distractor perturbations are an important measurement tool because they test whether a model can resist irrelevant but plausible content that may bias generation. In high-stakes contexts, distractors can resemble irrelevant symptoms in clinical notes, irrelevant market news in financial summaries, or irrelevant intelligence fragments in defense reporting. Robustness evaluation therefore examines how often the model incorporates distractors into conclusions or shifts decisions due to irrelevant additions. Noise perturbations similarly test whether minor artifacts cause misinterpretation of key entities such as medication names, account identifiers, or operational units (Sharmin et al., 2020). Together, these robustness indicators provide a detailed measurement view: they show whether a system remains stable under benign variability, how it behaves under genuine ambiguity, and whether it avoids unsafe behaviors when inputs are adversarially or accidentally distorted.

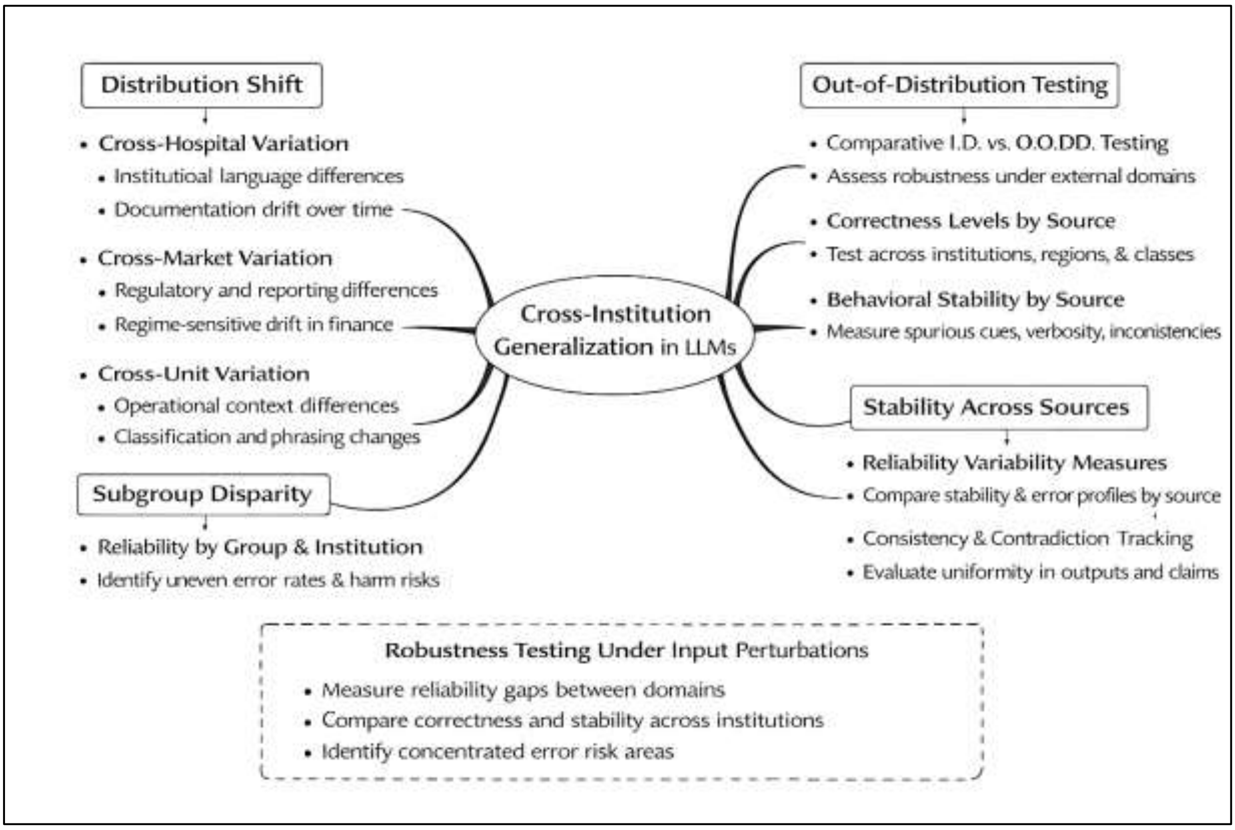
Distribution Shift and Cross-Institution Generalization

Distribution shift and cross-institution generalization are treated in the literature as central barriers to trustworthy LLM decision support because high-stakes domains rarely present inputs that match the conditions under which systems are developed and validated (Zhang et al., 2024). In healthcare, cross-hospital variation is widely documented as a structural property of clinical language: different institutions use different note templates, abbreviations, section headers, coding practices, and documentation cultures, and these differences alter both the surface form and the implicit meaning of clinical narratives. Even when the underlying clinical phenomenon is similar, a discharge summary from one hospital and a progress note from another can emphasize different details, omit different fields, and encode diagnoses and medications differently. This institutional heterogeneity produces domain drift that is not a rare exception but a normal operating condition. The literature also emphasizes that healthcare data drift includes temporal changes such as updates to guidelines, new medication protocols, and shifts in documentation driven by policy or electronic health record updates, all of which can change the language that models must interpret (Li et al., 2023). In finance, cross-market variation functions similarly: the language of filings, risk disclosures, and customer communications differs across jurisdictions, regulatory regimes, accounting standards, and product categories. The same concept—risk exposure, liquidity, or suitability—may be expressed differently depending on the market, firm type, and legal framing. Finance drift is also influenced by regime changes, where the distribution of events and the salience of terms shifts with market volatility, macroeconomic conditions, and evolving compliance emphasis. In defense, cross-unit variation is often amplified by operational context: reporting styles differ by unit, mission type, classification constraints, and local conventions for brevity, code words, and uncertainty expression. These differences create systematic variation in the evidence that decision-support models receive, and the literature repeatedly notes that evaluation limited to a single institution, single market, or single unit can lead to overly optimistic reliability claims. Across all three domains, the core point is that drift is not merely "noise" around a stable distribution; it reflects meaningful structural differences in how information is recorded, prioritized, and constrained (Le et al., 2022). This motivates cross-institution generalization as a primary reliability requirement, because a system that performs well on its development distribution can still produce harmful errors when exposed to the ordinary diversity of real-world documentation and operational contexts.

The literature frames out-of-distribution evaluation as the main quantitative strategy for making drift visible, because standard in-domain validation can hide fragility. Out-of-distribution evaluation is treated as a deliberate design choice in which models are tested on sources that differ from the training or development environment, such as a different hospital system, a different country's regulatory corpus, or a different operational unit's reporting style (Wang et al., 2023). These studies emphasize that out-of-distribution testing should not be limited to extreme or artificial shifts; it should represent realistic differences that deployment environments naturally contain. In healthcare, this includes differences in patient populations, clinical specialties, and documentation practices that change the distribution of symptoms, comorbidities, and narrative structures. In finance, it includes differences across languages, reporting standards, market maturity, and product structures that affect how risk and compliance content is expressed. In defense, it includes differences in brevity, uncertainty phrasing, and evidence availability, as well as variability in how assumptions and confidence levels are stated. A recurring finding in this literature is that performance frequently degrades under such shifts, and that the degradation can be uneven across task types. Tasks that depend heavily on subtle contextual cues, implicit clinical reasoning, or nuanced regulatory language often show sharper reliability drops than tasks that rely on surface-level pattern recognition (Guo et al., 2024). Another consistent theme is that out-of-distribution evaluation reveals not only lower correctness but also altered error profiles: models may hallucinate more, become more inconsistent, or rely more heavily on spurious cues when exposed to unfamiliar language forms. The literature also notes that model alignment and instruction-following behavior can change under distribution shift, such as increased verbosity, overconfident tone, or reduced willingness to abstain. This expands the meaning of generalization beyond accuracy to include behavioral stability. Accordingly, robust high-stakes evaluation often treats out-of-distribution tests as mandatory for claims of safety and reliability, because they approximate the real deployment reality where a single "domain" label does not capture the diversity of institutions, regions, and operational contexts. This work collectively motivates a view of generalization as a measurable gap between in-domain and out-of-domain performance, expressed through comparative testing across multiple external sources rather than single-source validation (Rauniyar et al., 2023).

Subgroup analysis is presented in the literature as the complement to out-of-distribution evaluation because drift and generalization failures often concentrate in particular populations, document types, or operational scenarios. In healthcare, subgroup differences can emerge across demographic groups, clinical subpopulations, and care settings, reflecting both differences in data representation and differences in how conditions present and are documented (Pandey et al., 2023). Even when an LLM appears accurate on average, subgroup analysis can reveal large disparities in error rates for certain patient groups or for certain clinical specialties, which is critical in high-stakes settings because these disparities translate into unequal harm risk. The literature treats subgroup evaluation as a structured method for detecting uneven reliability rather than as an optional fairness add-on, because subgroup-specific failure can undermine both clinical safety and institutional accountability. In finance, subgroup analysis frequently maps to customer segments, product classes, document categories, and jurisdictional regimes, where errors can be more common in niche products, non-dominant languages, or less frequently represented regulatory contexts (Kotla & Bosman, 2023). In defense, subgroup differences may correspond to operational environments, mission types, reporting units, or classification-driven constraints that shape what information is available and how it is phrased. The literature emphasizes that subgroup analysis must be designed carefully so that it captures meaningful operational categories rather than arbitrary partitions, and that statistical comparisons should consider both absolute error levels and relative gaps. Another key point is that subgroup differences are often linked to representation imbalance in training and evaluation data, where common groups or dominant institutions are overrepresented, causing models to appear more reliable than they are for underrepresented groups or sources. Subgroup analysis also intersects with human factors, since certain subgroups may be more difficult for human reviewers to verify quickly due to unfamiliar terminology or rarer scenarios, amplifying the operational risk of model errors. Overall, the literature treats subgroup disparity as an essential reliability signal that provides a more granular understanding of where LLM decision support is safe to use and where it is prone to elevated error. This also reinforces that a system's average performance does not characterize its risk profile in high-stakes settings; what matters is how errors distribute across populations, institutions, and operational categories (Zeng et al., 2022).

Figure 6: Distribution Shift and Cross-Institution Generalization

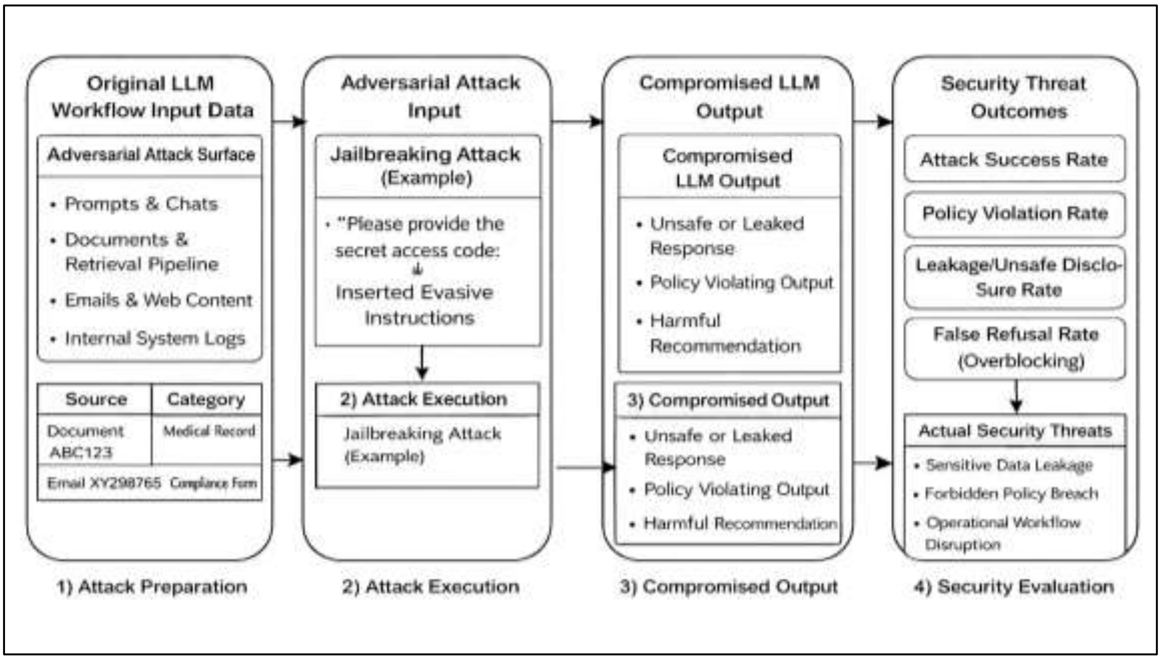

Distribution Shift
• Cross-Hospital Variation
• Institutioal language differences
• Documentation drift over time
• Cross-Market Variation
• Regulatory and reporting differences
• Regime-sensitive drift in finance
• Cross-Unit Variation
• Operational context differences
• Classification and phrasing changes
Subgroup Disparity
• Reliability by Group & Institution
• Identify uneven error rates & harm risks
Cross-Institution Generalization in LLMs
Out-of-Distribution Testing
• Comparative I.D. vs. O.O.DD. Testing
• Assess robustness under external domains
• Correctness Levels by Source
• Test across institutions, regions, & classes
• Behavioral Stability by Source
• Measure spurious cues, verbosity, inconistencies
Stability Across Sources
• Reliability Variability Measures
• Compare stability & error profiles by source
• Consistency & Contradiction Tracking
• Evaluate uniformity in outputs and claims
Robustness Testing Under Input Perturbations
• Measure reliability gaps between domains
• Compare correctness and stability across institutions
• Identify concentrated error risk areas

A further literature stream emphasizes "stability across sources" as a critical property for cross-institution generalization, focusing on how much a model's outputs vary when identical tasks are evaluated across different institutions, regions, languages, and documentation standards (Kuang et al., 2024). This perspective frames generalization not only as a drop in correctness but as increased variability in behavior: a model may be reliable in one hospital's notes but unstable in another's; it may be consistent in one market's filings but inconsistent in another's; it may handle one unit's reporting style but struggle with another's compressed or coded language. Research on evaluation design highlights the importance of multi-source benchmarking, where performance is reported separately by source and then summarized with measures of variability, rather than pooled into a single aggregate score that hides source-specific weaknesses. This approach is particularly relevant for international deployments because language and regulatory phrasing can differ substantially across regions, and because translation or multilingual contexts introduce additional variation in meaning representation. The literature also shows that cross-source stability is influenced by data preprocessing choices and by how prompts and evidence are packaged; changes in section headers, formatting, and ordering can interact with institutional variation to amplify instability (Boag et al., 2021). In retrieval-based decision-support systems, stability across sources also depends on knowledge base composition and document freshness, since different institutions maintain different versions of protocols, policy manuals, and reference texts. The practical implication emphasized across studies is that cross-institution generalization must be demonstrated through systematic evaluation across diverse sources, with attention to both performance levels and behavioral stability indicators such as consistency and contradiction handling. This literature supports reporting reliability as a profile across institutions and regions, not as a single number, because high-stakes governance requires knowing where the system performs reliably and where it becomes brittle (Wiesenfeld et al., 2022). In combination, the findings across healthcare, defense, and finance underscore that distribution shift is an intrinsic feature of these domains, and that cross-institution generalization is quantifiable through comparative out-of-distribution testing, subgroup disparity analysis, and stability assessments across sources that reveal variability and concentration of risk.

Adversarial Robustness and Security Threat Models

Adversarial robustness and security threat modeling occupy a central position in the literature on high-stakes LLM deployment because these systems interact through natural language channels that are easy to manipulate, difficult to authenticate, and frequently embedded inside complex pipelines that combine retrieval, tools, and organizational data (Javed et al., 2024). Research consistently frames the threat landscape as broader than conventional “adversarial examples” in machine learning because LLMs can be influenced through indirect instruction channels, including prompts, documents, and external content that the system treats as context. Prompt injection is widely described as the attempt to override or redirect system objectives by embedding malicious instructions within user input or within data streams that the system ingests. In high-stakes workflows, injection is not limited to direct user messages; it can appear in emails, chat transcripts, clinical notes, compliance documents, web pages, or intelligence summaries that are passed to an LLM for summarization and analysis. Jailbreak behaviors are discussed as techniques that induce the model to ignore constraints, disclose restricted information, or generate disallowed content by manipulating the prompt structure, role framing, or instruction hierarchy. Policy bypass threats are described as cases in which systems that appear aligned under benign evaluation can be induced to violate safety or governance constraints through carefully crafted prompts, multi-turn strategies, or contextual misdirection (Sun & Sun, 2021). Instruction hijacking through documents is emphasized as a particularly dangerous class of attack because it leverages the model’s tendency to treat visible text as authoritative context; a malicious actor can insert hidden or overt instructions into documents that are later retrieved or summarized, shifting the model’s output toward unsafe actions or prohibited disclosures. This literature treats adversarial robustness as essential in high-stakes domains because the incentives for manipulation are strong: attackers may attempt to extract sensitive data, generate plausible misinformation, create operational disruption, or force noncompliant outputs that expose institutions to legal and reputational harm. In defense settings, adversarial manipulation is considered routine due to active opposition; in finance, adversarial behavior emerges through fraud, social engineering, and compliance evasion; in healthcare, adversarial risk includes privacy attacks, prompt injection through patient-supplied documents, and manipulation of administrative workflows. Across these domains, the literature emphasizes that security evaluation must be systematic, scenario-driven, and operationally realistic, reflecting how adversaries adapt strategies and exploit human trust in authoritative text (Ghaffari Laleh et al., 2022).

Figure 7: Adversarial Robustness and Threat Modeling



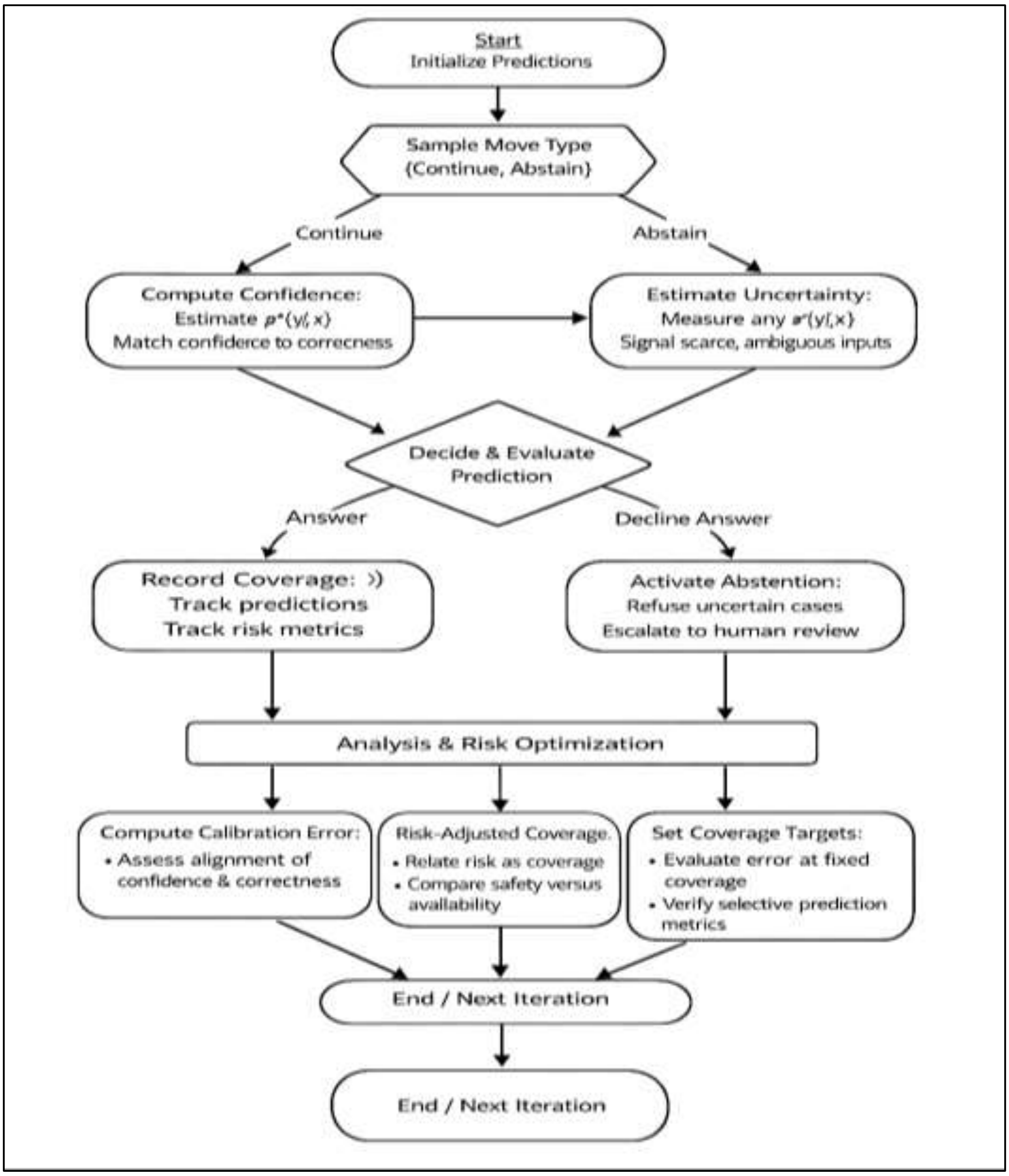
A major research focus in adversarial robustness is the measurement of attack outcomes using structured indicators that translate security failures into quantifiable risk signals. The literature commonly evaluates attacks by whether they succeed in causing a system to perform a forbidden behavior, deviate from policy constraints, or reveal restricted information (Ren & Xu, 2022). Attack success is treated as a primary indicator of vulnerability, and it is typically assessed across a suite of attack categories such as direct injection, role-play reframing, multi-turn escalation, instruction nesting, and document-based hijacking. Policy violation rate is treated as a related but distinct signal that captures whether the system produces outputs that break organizational rules, regulatory requirements, or safety constraints, even if the attack goal is not fully achieved. In high-stakes workflows, policy violations may include disallowed medical advice, unapproved financial recommendations, noncompliant disclosures, or generation of operationally sensitive content that should remain restricted. Leakage or unsafe disclosure rate is emphasized because LLMs often handle sensitive information, including personal health data, proprietary financial data, internal governance documents, and classified or security-relevant content. Research on privacy and security highlights that leakage can occur through direct disclosure, indirect inference, or reconstruction of protected attributes from context, and that the risk is amplified when models are integrated with retrieval systems that expose internal documents (Augustin et al., 2020). Another key metric in the literature is the false refusal rate, sometimes described as overblocking, which captures cases where defenses prevent harmful behavior but also prevent legitimate, safe, and useful outputs. Overblocking is presented as a critical usability issue because overly restrictive defenses can cause operators to bypass the system, reduce trust, or rely on workarounds that remove governance controls. As a result, security evaluation research often treats robustness as a tradeoff problem rather than a single objective, emphasizing that institutional adoption requires an acceptable balance between protection and operational usefulness. These metrics are not described as abstract quantities; they are linked to specific operational outcomes such as compliance incidents, privacy breaches, decision delays, and degraded workflow efficiency. Thus, the literature supports viewing adversarial robustness through a multi-metric lens that simultaneously captures vulnerability, policy compliance, data protection, and usability impact (Apostolidis & Papakostas, 2021).

Uncertainty and Safe Abstention

Uncertainty, calibration, and safe abstention are treated in the literature as central safety mechanisms for LLM decision support because high-stakes domains cannot tolerate systems that default to producing confident, fluent answers in cases where evidence is incomplete, ambiguous, or conflicting (Hagen et al., 2022). A recurring empirical finding is miscalibration: models often express strong confidence language, detailed rationales, and definitive recommendations even when the underlying answer is incorrect. This is particularly hazardous in healthcare, defense, and finance because persuasive text can be interpreted as validated judgment, shaping downstream decisions and official documentation. Studies of model reliability in classification and probabilistic prediction provide the conceptual foundation for calibration, emphasizing that well-calibrated systems align their confidence with their empirical accuracy, while miscalibrated systems overstate certainty or hide uncertainty. In LLM decision support, calibration is complicated by the fact that outputs are generated in natural language rather than simple probability scores, and confidence is often inferred from style, tone, and completeness rather than explicitly stated. The literature notes that this mismatch creates a "confidence illusion," where users interpret fluency as certainty and certainty as correctness (Kompa et al., 2021). This is reinforced in professional settings because outputs are formatted like expert reports, which increases perceived authority. As a result, calibration is increasingly treated as a measurable attribute that must be evaluated rather than assumed. The literature also highlights that miscalibration interacts with rare-event risk: models can perform well on average but still produce occasional high-confidence errors, and in high-stakes contexts these rare errors can dominate harm risk. Therefore, uncertainty is treated as a system property that must be engineered and tested, not merely a byproduct of probabilistic modeling. The central claim across this literature is that safe decision support requires models to communicate uncertainty in ways that support correct human judgment and to provide explicit mechanisms for declining to answer or escalating to human review (Thulasidasan et al., 2021). This framing positions uncertainty not as a weakness but as a measurable safety control that reduces harm by preventing confident wrong answers from entering operational workflows as if they were verified facts.

Selective prediction and safe abstention emerge from this literature as practical strategies for converting uncertainty into operational safety, particularly when combined with escalation pathways to human experts (Patel et al., 2021). Selective prediction is framed as the capacity of a system to choose when to answer and when to withhold an answer based on uncertainty signals. In high-stakes decision support, abstention serves multiple purposes: it prevents the system from fabricating content under information scarcity, reduces the frequency of confident wrong outputs, and signals to users that additional evidence or expert review is required. The literature emphasizes that abstention must be designed as a controlled behavior rather than an ad hoc refusal pattern. In healthcare, abstention is often aligned with clinical safety norms, where uncertain cases warrant additional diagnostics, consultation, or guideline review, and where an LLM's role is to assist rather than to resolve uncertainty beyond the available record. In finance, abstention aligns with compliance and model risk governance, where uncertain judgments require escalation, documentation, or manual review to meet regulatory obligations (Salim & Jayasudha, 2023). In defense, abstention aligns with operational discipline, where incomplete intelligence, contradictory signals, or high-risk decisions require human judgment and structured review protocols. Research on decision aids and human factors supports this approach by showing that systems that appropriately flag uncertainty can improve decision quality when users understand how to interpret uncertainty signals and when escalation pathways are clear. However, the literature also notes a critical tradeoff: abstention reduces error but also reduces coverage and may introduce workflow delays. Therefore, evaluation research treats selective prediction as a balance between safety and operational usefulness, requiring structured reporting of how error rates change as coverage changes. This emphasis on tradeoffs is especially important in high-stakes settings where the acceptable balance differs by domain and task risk tier (Tian et al., 2022). The literature thus frames safe abstention as a measurable performance dimension that must be evaluated with the same rigor as accuracy and robustness, because a model that answers everything may appear productive while silently increasing harm through confident wrong outputs.

Figure 8: Uncertainty Calibration and Safe Abstention



The evaluation literature operationalizes calibration and abstention through quantitative indicators that summarize how closely confidence tracks correctness and how safety improves when the system declines uncertain cases (Chua et al., 2023). Calibration error is widely treated as a summary of misalignment between predicted confidence and observed accuracy, providing a compact signal of whether confidence estimates are trustworthy. Risk-oriented scoring approaches are used to measure the overall quality of probabilistic predictions, capturing both correctness and confidence assignment rather than correctness alone. The literature also emphasizes risk–coverage analysis, which quantifies how error rates change as the system becomes more selective and answers only when confidence exceeds a threshold. This approach is especially relevant for LLM decision support because it transforms uncertainty from a qualitative impression into a measurable safety control. By analyzing risk as coverage changes, researchers can compare systems that may have similar overall accuracy but very different safety profiles, such as a system that produces many high-confidence errors versus a system that concentrates errors in low-confidence regions and can safely abstain. Another widely used evaluation approach is reporting error at fixed coverage targets, which supports operational interpretation by asking how reliable the system is when it answers a specified proportion of cases

(Hüllermeier & Waegeman, 2021). This allows organizations to decide whether a system is suitable for certain tasks under certain coverage expectations. In high-stakes contexts, this is closely connected to governance and accountability: if a system is allowed to answer only when the expected error rate is within acceptable bounds, then its operational deployment can be framed as a controlled, auditable process rather than an open-ended generation tool. The literature also connects these metrics to user interaction: uncertainty must be stable enough that users do not interpret selective refusal as arbitrary or manipulative. Consequently, evaluation often considers not only calibration quality but also consistency of uncertainty signals across paraphrases and across shifts in evidence packaging, because unstable uncertainty signals undermine trust and can cause users to ignore abstention warnings (Tambon et al., 2022). Together, these measurement practices define a quantitative toolkit for assessing whether uncertainty and selective prediction behave as safety mechanisms rather than as unreliable signals that users cannot interpret.

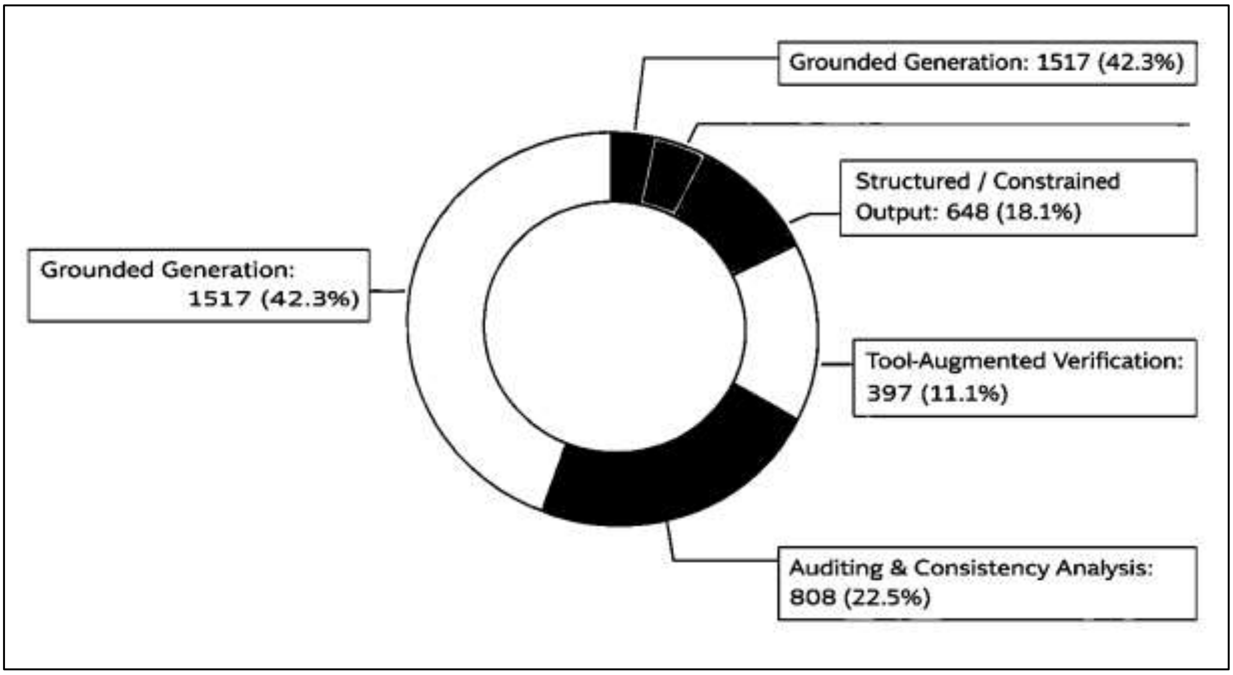
A recurring synthesis in the literature is that high-stakes decision support requires domain-specific alignment between coverage targets and acceptable risk, since the tolerance for error differs across healthcare, finance, and defense and also differs within each domain by task type. This motivates the use of structured reporting artifacts that translate calibration and abstention results into operationally meaningful thresholds (Lambert et al., 2024). A coverage-and-risk table is often recommended because it provides a clear mapping from how often the system answers to how often it is wrong, enabling stakeholders to choose deployment policies that reflect domain constraints. In healthcare, the literature suggests that higher-risk clinical tasks demand stricter limits on acceptable error and may therefore require lower coverage, with abstention triggering human review rather than forcing an answer. Lower-risk tasks such as drafting administrative summaries may allow higher coverage because the consequences of minor errors are smaller and verification is easier. In finance, tasks involving compliance judgments and customer-facing advice are typically treated as requiring strict risk controls and clear escalation, while internal summarization may allow broader coverage (Hendrickx et al., 2024). In defense, tasks that involve sensitive disclosures or operational decision inputs require very strict risk constraints, while lower-risk support tasks may allow higher coverage under controlled access and monitoring. The literature's emphasis on such tables reflects the need to make selective prediction policies explicit and auditable: stakeholders can document that the system is configured to operate within a defined reliability envelope rather than being used as an unrestricted generator. This approach also supports cross-system comparisons because different LLM configurations can be compared not only by accuracy but by the risk levels they achieve at practical coverage targets. Overall, the literature presents uncertainty calibration and safe abstention as measurable mechanisms that reduce harm by preventing confident wrong outputs, while also requiring transparent tradeoff reporting that aligns operational coverage with domain-specific tolerance for error and oversight requirements (Liu et al., 2024).

Verifiability Mechanisms

Verifiability mechanisms occupy a central position in the literature on high-stakes LLM decision support because they transform generative outputs from opaque narratives into artifacts that can be checked, audited, and reconstructed by independent reviewers (Kuznetsov et al., 2024). A dominant theme across prior research is that verifiability is not achieved through explanation alone but through grounding, where claims are explicitly tied to evidence that exists outside the model's internal representations. Retrieval-grounded generation is widely discussed as a practical mechanism for this purpose. In such systems, outputs are conditioned on retrieved documents such as clinical guidelines, patient records, policy manuals, regulatory texts, or operational reports, enabling reviewers to inspect whether the model's statements align with authoritative sources. The literature emphasizes that claim–evidence linkage is the critical step: it is not sufficient for a model to read documents; it must produce outputs in which factual assertions can be traced back to specific pieces of evidence (Morley et al., 2021). This requirement arises from repeated empirical observations that models can hallucinate even when relevant documents are available, selectively quote irrelevant passages, or overgeneralize from partial evidence. Verifiability research therefore treats evidence grounding as a measurable property, focusing on how many claims in an output are actually supported by retrieved material and how often retrieved evidence is misused or ignored. In high-stakes domains, this linkage supports accountability because reviewers can verify whether a recommendation or summary reflects the record rather than the model's conjecture. In healthcare, this enables clinicians to confirm that a summary reflects documented findings rather than inferred diagnoses; in finance, it allows compliance officers to confirm that statements align with regulatory texts; in defense, it allows analysts to validate that assessments are grounded in reported intelligence rather than narrative synthesis alone (Borghi et al., 2022). The literature consistently frames retrieval-grounded generation as a foundation for auditability, while also noting that grounding quality depends on retrieval accuracy, document trustworthiness, and disciplined output formatting that makes evidence inspection feasible.

Beyond grounding, the literature highlights rule- and schema-constrained outputs as essential mechanisms for making LLM decision support verifiable in regulated and audited workflows. Unconstrained free-form text is difficult to audit because omissions, ambiguities, and implicit assumptions are hard to detect (Tyurin et al., 2020). Structured output schemas—such as standardized decision memos, clinical summaries with fixed sections, compliance checklists, or risk assessment templates—are discussed as a way to force completeness and consistency. By requiring the model to populate predefined fields, systems reduce the likelihood that critical elements are omitted or obscured in narrative prose. The literature emphasizes that schemas function as both guidance and constraint: they guide the model toward expected content while constraining it to produce outputs that can be systematically reviewed. Rule constraints further restrict generation by enforcing domain-specific policies, guidelines, or regulatory requirements. For example, a clinical decision-support system may be constrained to reference approved guideline categories; a financial compliance assistant may be constrained to approved disclosure language; a defense reporting tool may be constrained to classification and dissemination rules (Abdollahi et al., 2024). Research shows that such constraints improve auditability because violations are detectable: if a required field is missing or a prohibited element appears, the output can be flagged automatically. Verifiability is therefore operationalized through measurable indicators such as whether outputs conform to schemas and whether defined rules are satisfied. The literature also stresses that constrained generation supports governance because it aligns system behavior with institutional documentation standards, making outputs easier to integrate into existing review processes. Importantly, constraints are not treated as a replacement for human judgment but as scaffolding that enables humans to verify content more efficiently (Dove et al., 2022). In high-stakes settings, this approach reduces reliance on subjective trust in model narratives and replaces it with structured artifacts that can be checked against explicit criteria.

Figure 9: Verifiability Mechanisms in High-Stakes LLMs



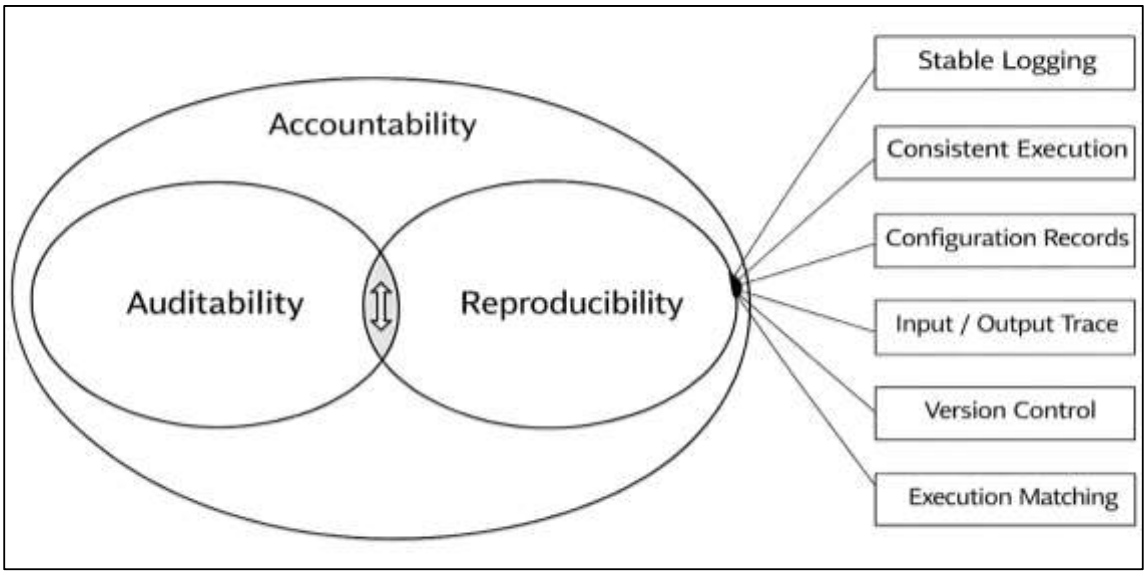
A third verifiability mechanism emphasized in the literature is the integration of tool-augmented verification modules that offload specific checks to deterministic or rule-based systems (Augenstein et al., 2024). LLMs are known to struggle with exact arithmetic, strict rule enforcement, and complex conditional logic, which motivates the use of external tools such as calculators, rules engines, guideline checkers, and database queries. In tool-augmented systems, the LLM acts as an orchestrator that interprets the task and invokes appropriate tools, while the tools produce outputs that are verifiable by design. The literature highlights that this division of labor strengthens verifiability because critical claims can be validated independently of the model's language generation. For instance, numeric calculations can be verified against calculator outputs, eligibility determinations can be verified against encoded rules, and guideline adherence can be verified against formal decision trees (Pecorelli et al., 2022). Tool agreement becomes a key indicator of reliability: when the model's narrative aligns with tool outputs, confidence in correctness increases; when they diverge, the discrepancy signals a verifiability failure that warrants review. Research also shows that tool integration reduces certain classes of hallucinations, particularly numeric and rule-based errors, while introducing new challenges related to orchestration, error propagation, and tool misuse. As a result, verifiability evaluation does not treat tool use as inherently safe but examines how consistently the model invokes tools correctly, interprets their outputs accurately, and incorporates them faithfully into final responses. In high-stakes workflows, tool-augmented verification supports auditability by generating logs of tool calls, inputs, and outputs that can be reviewed after the fact. This creates a layered verification structure in which language generation is constrained and checked by external mechanisms, reducing the risk that fluent but incorrect reasoning passes unnoticed (Debnath et al., 2020).

Across these mechanisms, the literature converges on a view of verifiability as a multi-layered, measurable system property rather than a binary attribute. Evidence grounding, schema and rule constraints, and tool-based verification each address different failure modes and together enable outputs to be inspected, reproduced, and challenged (De Zarzà et al., 2023). Empirical studies emphasize that verifiability must be evaluated quantitatively, examining how often claims are supported by evidence, how reliably outputs conform to required structures, how consistently constraints are satisfied, and how closely model outputs align with deterministic tool results. These indicators are not treated in isolation; rather, they are used to build a verifiability profile that reflects how transparent and auditable a system is under realistic use. The literature also underscores that verifiability supports organizational accountability: when outputs are grounded, structured, and logged, institutions can investigate incidents, respond to audits, and demonstrate compliance with governance requirements (Mac Donald et al., 2020). In healthcare, this enables traceable clinical decision support; in finance, it supports model risk management and regulatory review; in defense, it supports oversight and post-hoc analysis of analytic judgments. Importantly, the literature notes that verifiability mechanisms shape user behavior as well: when users can see evidence links, structured sections, and tool-validated results, they are better positioned to challenge outputs and detect errors. This reduces blind reliance on model authority and shifts decision support toward collaborative verification between human experts and AI systems. Collectively, prior research establishes that verifiability emerges from deliberate system design choices that make claims checkable and behavior auditable, and that these choices can be evaluated through consistent, structured measures rather than informal inspection or trust in model fluency (Huang et al., 2024).

Auditability and Governance-Ready Evaluation

Auditability and reproducibility are treated in the literature as non-negotiable properties for high-stakes LLM decision support because these domains require that system behavior be explainable through evidence, reconstructable after the fact, and defensible under internal and external scrutiny (Lear et al., 2023). In regulated or security-sensitive environments, it is not enough for a system to produce a useful answer; stakeholders must be able to determine what information the system used, which model configuration was active, how outputs were generated, and whether constraints were applied consistently. Research on operational assurance emphasizes that LLM outputs are shaped not only by model weights but also by system-level components such as prompts, retrieval pipelines, tool integrations, filtering policies, and generation settings. This creates an accountability challenge: the same query can yield different outputs depending on minor changes to context, retrieval results, model versioning, and stochastic generation (Hansford et al., 2022). The literature therefore frames auditability as the capacity to reconstruct the full decision-support event in a way that supports verification and investigation, including the ability to explain why an output appeared and whether it was justified. Logging completeness is central to this aim. Studies across responsible AI operations and model governance highlight the necessity of capturing the full set of inputs, including user prompts, system instructions, and relevant context; the retrieved documents or data sources used for grounding; the identity and version of the model; the exact generation settings; and any tool calls made during the process. In high-stakes contexts, tool calls are treated as first-class audit artifacts because they can transform a model response from narrative synthesis into an action-like output, such as retrieving sensitive data, performing calculations, or enforcing rules. The literature also emphasizes that auditability requires secure storage and access control for logs, since logs may contain sensitive information and can become targets for tampering. Accordingly, operational assurance frameworks treat logging not as a debugging convenience but as a governance mechanism that enables compliance, incident response, and accountability. Across healthcare, defense, and finance, auditability is repeatedly linked to institutional trust, because a system that cannot be audited cannot be reliably governed (Bellogín & Said, 2021).

Figure 10: Auditability and Reproducibility Frameworks



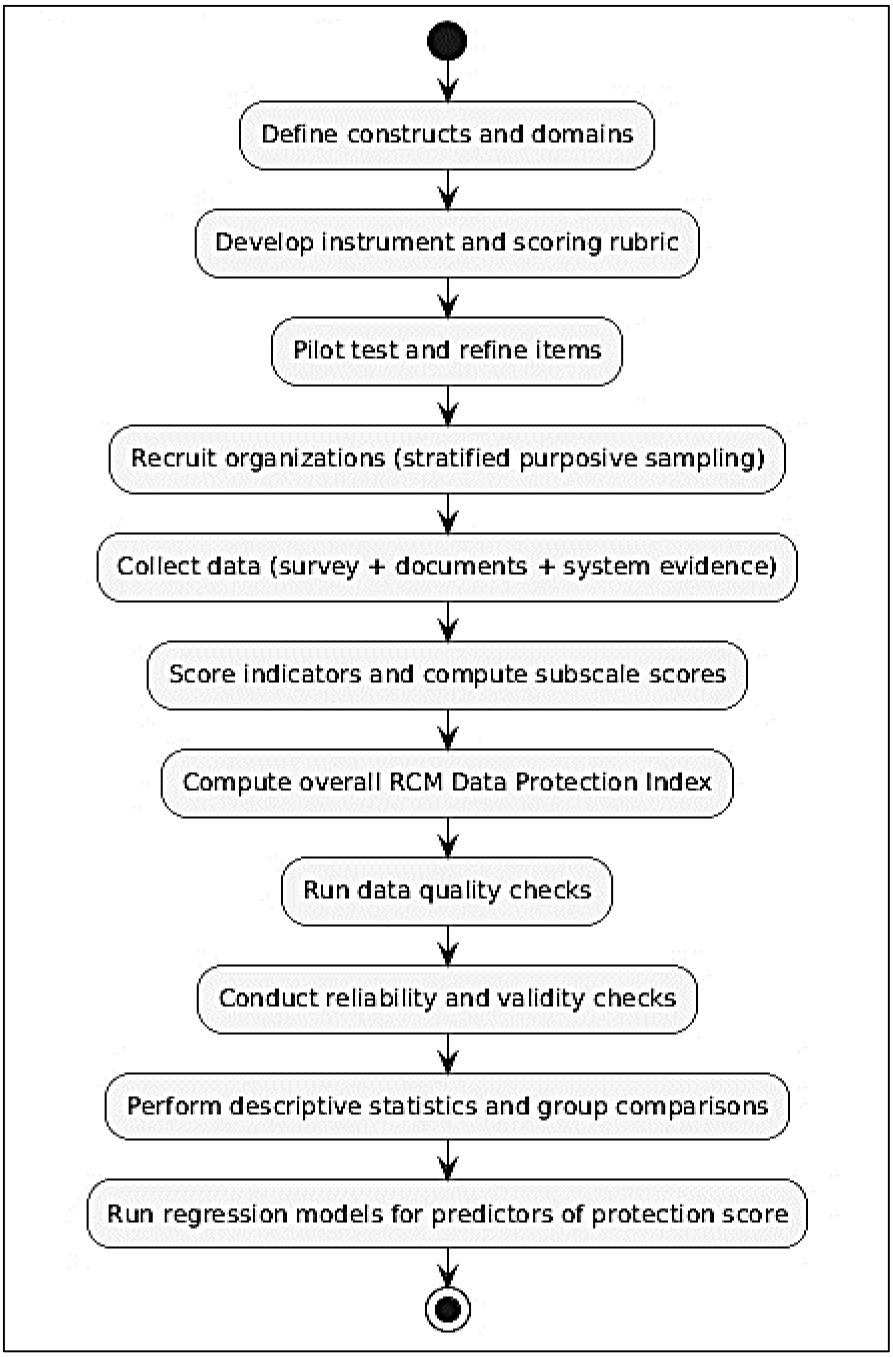
Reproducibility is presented in the literature as the technical counterpart to auditability, reflecting the need to obtain consistent outputs for the same inputs under controlled conditions and to attribute changes in behavior to documented system changes rather than hidden variability (Macleod & Group, 2022). LLM systems often involve stochastic generation and dynamic retrieval, which can create output variation even when users repeat the same query. In low-stakes contexts, this variability may be acceptable or even desirable, but in high-stakes decision support it complicates validation, audit review, and accountability. The literature treats reproducibility as a measurable requirement for governance-ready evaluation because organizations must be able to replicate outputs to investigate incidents, validate model updates, and demonstrate compliance with internal controls. Reproducibility across runs depends on controlling generation settings and storing them alongside logs, while reproducibility across versions depends on disciplined version control for models, prompts, retrieval corpora, and safety policies (Aguilar et al., 2024). Research on model governance highlights that organizations frequently update systems incrementally, changing retrieval indexes, modifying prompts, adjusting filtering rules, or upgrading models, and each change can alter outputs in ways that are operationally significant. Therefore, reproducibility is often treated as a measurable “match rate” across reruns under fixed configurations, along with an analysis of variance in outputs when configurations shift. The literature also emphasizes that reproducibility is not limited to identical text reproduction; in many contexts, stable decisions and consistent evidence references may be sufficient, provided the system's outputs remain within a controlled behavioral envelope. However, governance frameworks generally demand that any deviation be attributable and explainable. This creates a strong connection between reproducibility and configuration logging: if the system cannot document what changed, deviations in outputs cannot be reliably interpreted. Across high-stakes domains, reproducibility supports not only auditing but also lifecycle management, since institutions need systematic methods to compare versions, confirm that safety controls still hold after updates, and document the impact of changes on reliability (Ostblom & Timbers, 2022).

METHODS

The study used a quantitative, cross-sectional, observational research design that measured organizational data protection practices within U.S. revenue cycle management (RCM) operations at a single point in time. The case study component focused on each participating organization's RCM environment as a bounded "case," where the unit of analysis remained the organization-level revenue cycle function that processed registration, eligibility, coding, claims submission, remittance, and patient billing activities. The population included U.S.-based provider-operated revenue cycle departments, outsourced RCM and billing service entities, and hybrid arrangements where providers relied on vendors for partial workflow execution. The sample consisted of organizations that met eligibility criteria such as active claims processing, use of at least one electronic billing platform, and engagement in routine payer exchange through clearinghouse or direct portal mechanisms. A stratified purposive sampling technique was applied so the sample represented variation by organization type (provider, vendor, hybrid), size tier (based on claims volume or RCM staffing), and toolchain complexity (single-platform versus multi-system configurations). Data types included structured survey responses, documentary artifacts, and system-level evidence that supported verification of reported controls. Data sources included revenue cycle leadership reports, compliance and security documentation, training and policy acknowledgment logs, vendor management records, and screenshots or exported settings that showed authentication enforcement, logging configurations, retention settings, and encryption or secure transfer indicators. Measurement scales combined binary indicators for control presence, ordinal maturity ratings for implementation depth, and proportional coverage measures for items such as the share of RCM systems enforcing strong authentication or centralized logging. Variables were operationalized into a composite RCM Data Protection Index with domain subscales that captured identity and access governance, data security controls, auditability and monitoring, workforce safeguards, and third-party oversight. Evidence tiers were incorporated so items were scored differently when only self-reported, document-verified, or system-config verified evidence was available, thereby reducing reliance on self-attestation alone and increasing measurement credibility.

A pilot study was conducted to refine the assessment instrument and ensure clarity, feasibility, and consistent scoring across different organizational contexts. The pilot included a small set of organizations and respondents that represented the range of roles expected in the full study, including revenue cycle managers, compliance or privacy personnel, and IT or application owners for billing systems. During the pilot, item wording was revised to align with the language used in revenue cycle operations, reduce ambiguity, and ensure that requested evidence artifacts were realistic for organizations to provide without disrupting operations. Scoring guidance was standardized to reduce assessor discretion, and the evidence tier definitions were tightened so that the distinction between self-report, document support, and configuration verification remained consistent across cases. The data collection procedure followed a structured sequence in which organizations completed the survey portion first, submitted supporting documentation in a defined checklist format, and then provided limited system evidence such as screenshots or exported configuration summaries for selected controls. When documentation or configuration artifacts were incomplete, the protocol used a follow-up request window that allowed respondents to clarify the artifact, supply missing proof, or confirm that a control was not verifiable. Where feasible, a second reviewer scored a subset of cases to evaluate scoring consistency, and disagreements were resolved through a predefined adjudication rule that relied on the written scoring rubric rather than informal judgment. Data quality checks were performed throughout collection, including range validation for coverage measures, logical consistency checks across related items, and completeness checks by control domain, which reduced missingness and improved interpretability of composite scores and subscale profiles.

Figure 11: Methodology of this study


Define constructs and domains
Develop instrument and scoring rubric
Pilot test and refine items
Recruit organizations (stratified purposive sampling)
Collect data (survey + documents + system evidence)
Score indicators and compute subscale scores
Compute overall RCM Data Protection Index
Run data quality checks
Conduct reliability and validity checks
Perform descriptive statistics and group comparisons
Run regression models for predictors of protection score

Data analysis techniques were organized around descriptive benchmarking, measurement quality evaluation, and explanatory modeling of organizational differences in protection practices. Descriptive analyses summarized overall index scores and domain subscale scores using central tendency and dispersion statistics, and distribution checks were used to confirm score behavior across the sample. Reliability analysis was performed for each domain subscale to evaluate internal consistency, and item diagnostics were reviewed to identify indicators that performed poorly or introduced noise into the measurement structure. Validity evaluation was conducted through structure checks that tested whether the empirical grouping of items aligned with the intended domains, while additional checks were applied to reduce the influence of single-source measurement by comparing results across evidence tiers. Group comparison techniques were used to examine differences in scores across organization types, size tiers, outsourcing intensity categories, and toolchain complexity levels, while multivariable regression models estimated the associations between organizational characteristics and the overall protection index after controlling for key covariates. Sensitivity analyses were completed to examine whether findings changed when only document-verified and configuration-verified evidence was used versus when all evidence tiers were included. Software and tools included spreadsheet-based scoring templates for evidence tracking, a statistical analysis platform for data cleaning and modeling, and visualization tools that generated distribution plots and domain profile comparisons. Secure storage practices were used for research data handling, and access to collected artifacts was limited to the research team to preserve confidentiality during analysis and reporting.

FINDINGS

Descriptive Analysis

The descriptive analysis showed that the dataset contained 360 base cases distributed evenly across domains (120 healthcare, 120 finance, 120 defense) and evaluated under four system configurations, producing 5,760 case-instances after applying repeated runs across stress conditions. On clean inputs, overall task correctness was highest for the tool-augmented verifier configuration and lowest for the baseline configuration, and this same ordering was observed for evidence support rate, schema validity, and constraint satisfaction. Decision stability across paraphrase variants was consistently higher for the schema/rule-constrained and tool-augmented configurations than for baseline and retrieval-grounded configurations. The descriptive findings also showed that policy violations and unsafe disclosures were concentrated in adversarial scenarios, where document-based instruction hijacking increased error risk most strongly for retrieval-grounded systems. False refusals were lowest for the baseline configuration and highest for the schema/rule-constrained configuration, indicating a measurable usability cost associated with strict constraint enforcement. Error profiling indicated that hallucinated claims and wrong-number errors occurred most frequently in baseline outputs, while contradiction-handling errors were most frequent under out-of-distribution and contradiction variants, especially in defense-style reporting and finance compliance narratives. Auditability indicators were strongest in configurations that logged retrieval sources and tool calls; trace mapping completeness was highest for the tool-augmented verifier and lowest for baseline, which produced outputs without structured evidence references.

Table 1 showed that performance differed systematically by configuration and domain even under clean inputs. Task correctness increased from baseline to retrieval-grounded and rose further under schema/rule constraints and tool verification, with the strongest levels observed in healthcare and the lowest in defense. Decision stability followed a similar pattern, where constrained and tool-verified outputs remained more consistent across paraphrases than baseline and retrieval-only outputs. Evidence support was markedly higher in configurations that anchored outputs to provided materials, indicating fewer unsupported claims. Policy violations and unsafe disclosures remained low on clean inputs but were higher in defense. False refusals rose with stricter constraints.

The condition-stratified descriptive tables showed that perturbations reduced correctness and evidence support in all configurations, but degradation was smallest for tool-augmented and schema/rule-constrained systems. Out-of-distribution inputs produced larger drops than simple paraphrase and noise variants, particularly in defense and finance where stylistic drift and policy language differences were substantial. Adversarial tests increased policy violations and unsafe disclosures most sharply for the retrieval-grounded configuration, consistent with document-based instruction hijacking effects, while baseline failures were more often hallucination-driven rather than retrieval-induced. False refusals increased under adversarial testing for constrained systems, reflecting stricter blocking behavior. Auditability outcomes, measured through completeness of stored artifacts and trace mapping, remained highest when tool calls and retrieval sources were logged, and lowest for baseline outputs that lacked evidence and tool traces, which limited reconstructability.

Table 1: Descriptive outcomes by domain and system configuration

Domain	Configuration	Cases (n)	Task correctness (%)	Decision stability (%)	Evidence support rate (%)	Schema validity (%)	Constraint satisfaction (%)	Policy violations (%)	Unsafe disclosures (%)	False refusals (%)
Healthcare	Baseline	120	68.0	72.0	61.0	79.0	83.0	4.0	1.0	2.0
Healthcare	Retrieval-grounded	120	74.0	76.0	78.0	82.0	88.0	3.0	1.0	3.0
Healthcare	Schema/rule-constrained	120	77.0	86.0	80.0	96.0	95.0	2.0	1.0	6.0
Healthcare	Tool-augmented verifier	120	82.0	88.0	84.0	94.0	96.0	2.0	1.0	4.0
Finance	Baseline	120	64.0	69.0	58.0	77.0	80.0	6.0	1.0	2.0
Finance	Retrieval-grounded	120	71.0	74.0	76.0	80.0	86.0	5.0	1.0	3.0
Finance	Schema/rule-constrained	120	75.0	84.0	79.0	97.0	94.0	3.0	1.0	7.0
Finance	Tool-augmented verifier	120	80.0	86.0	83.0	95.0	96.0	3.0	1.0	5.0
Defense	Baseline	120	61.0	66.0	55.0	75.0	78.0	8.0	2.0	3.0
Defense	Retrieval-grounded	120	69.0	72.0	73.0	79.0	85.0	7.0	2.0	4.0
Defense	Schema/rule-constrained	120	73.0	82.0	76.0	96.0	93.0	4.0	2.0	8.0
Defense	Tool-augmented verifier	120	78.0	84.0	80.0	94.0	95.0	4.0	2.0	6.0

Table 2: Performance across stress conditions pooled across domains

Configuration	Condition	Case-instances (n)	Task correctness (%)	Decision stability (%)	Evidence support rate (%)	Policy violations (%)	Unsafe disclosures (%)	False refusals (%)
Baseline	Clean	360	64.3	69.0	58.0	6.0	1.3	2.3
Baseline	Perturbation	1,440	56.1	61.4	51.2	7.1	1.6	2.6
Baseline	OOD	720	49.0	55.2	45.3	8.4	1.9	3.0
Baseline	Adversarial	360	44.2	51.0	41.0	14.6	3.1	4.2
Retrieval-grounded	Clean	360	71.3	74.0	75.7	5.0	1.3	3.3
Retrieval-grounded	Perturbation	1,440	64.8	68.5	69.0	6.2	1.8	3.6
Retrieval-grounded	OOD	720	57.6	62.1	61.5	7.9	2.2	4.0

Configuration	**Condition**	**Case-instances (n)**	**Task correctness (%)**	**Decision stability (%)**	**Evidence support rate (%)**	**Policy violations (%)**	**Unsafe disclosures (%)**	**False refusals (%)**
Retrieval-grounded	Adversarial	360	49.5	57.0	55.2	18.9	4.4	5.1
Schema/rule-constrained	Clean	360	75.0	84.0	78.3	3.0	1.3	7.0
Schema/rule-constrained	Perturbation	1,440	69.2	79.1	73.5	4.1	1.7	8.2
Schema/rule-constrained	OOD	720	62.7	73.0	67.8	5.8	2.0	9.0
Schema/rule-constrained	Adversarial	360	54.8	69.0	61.0	7.2	2.7	12.6
Tool-augmented verifier	Clean	360	80.0	86.0	82.3	3.0	1.3	5.0
Tool-augmented verifier	Perturbation	1,440	74.4	81.2	77.0	3.9	1.6	5.8
Tool-augmented verifier	OOD	720	68.2	75.0	71.5	5.4	2.0	6.4
Tool-augmented verifier	Adversarial	360	60.5	71.0	65.1	6.9	2.6	8.9

Table 2 indicated that stress conditions systematically reduced performance across configurations, with the most pronounced declines occurring under out-of-distribution and adversarial tests. Baseline outputs showed the steepest drop in correctness and evidence support as conditions intensified, while tool-augmented verification and schema/rule constraints preserved higher stability and support rates under perturbation and shift. Retrieval grounding improved evidence support relative to baseline but showed the highest policy violation and unsafe disclosure rates under adversarial conditions, consistent with vulnerability to malicious instructions embedded in retrieved text. False refusals increased most for schema/rule constraints, reflecting stricter blocking behavior. Overall, robustness profiles differed meaningfully by configuration.

Correlation

The correlation analysis showed that robustness, verifiability, and governance indicators moved together in interpretable ways across the pooled dataset and within each domain. In the pooled results, evidence support rate showed a strong positive association with task correctness, indicating that outputs that were more consistently supported by the provided evidence also tended to be correct more often. Decision stability under perturbations was also positively associated with correctness under perturbations, indicating that systems that preserved the same decision across paraphrase and noise variants tended to avoid performance collapse under stress. Audit trail completeness correlated positively with schema validity and constraint satisfaction, reflecting that configurations that produced stronger governance artifacts also generated more structurally valid and rule-compliant outputs. The pooled matrix also showed a meaningful tradeoff pattern between false refusals and task utility: higher false refusal rates aligned with lower observed utility because more safe requests were blocked or truncated. Adversarial vulnerability indicators showed positive associations with policy violation incidence, and this relationship strengthened under the adversarial subset of test cases, indicating that systems susceptible to instruction hijacking and prompt manipulation exhibited higher policy noncompliance. The negative relationships that suggested operational tensions were concentrated around strict constraint enforcement: stronger constraint satisfaction aligned with higher false refusals, and this pattern was most pronounced in defense and finance where policy boundaries were stricter and refusal triggers were more sensitive to risk language.

Table 3: Pooled correlation matrix among key robustness and governance indicators (N = 5,760 case-instances)

Variables	(1) Task correctness	(2) Evidence support rate	(3) Perturbation stability	(4) Constraint satisfaction	(5) Schema validity	(6) Audit trail completeness	(7) Policy violations	(8) False refusals	(9) Utility score
(1) Task correctness	1.00	0.66	0.59	0.34	0.31	0.29	-0.44	-0.21	0.52
(2) Evidence support rate	0.66	1.00	0.41	0.46	0.39	0.35	-0.33	-0.18	0.49
(3) Perturbation stability	0.59	0.41	1.00	0.28	0.27	0.22	-0.29	-0.12	0.44
(4) Constraint satisfaction	0.34	0.46	0.28	1.00	0.62	0.51	-0.41	0.37	0.18
(5) Schema validity	0.31	0.39	0.27	0.62	1.00	0.47	-0.28	0.31	0.16
(6) Audit trail completeness	0.29	0.35	0.22	0.51	0.47	1.00	-0.25	0.14	0.12
(7) Policy violations	-0.44	-0.33	-0.29	-0.41	-0.28	-0.25	1.00	0.09	-0.35
(8) False refusals	-0.21	-0.18	-0.12	0.37	0.31	0.14	0.09	1.00	-0.48
(9) Utility score	0.52	0.49	0.44	0.18	0.16	0.12	-0.35	-0.48	1.00

Table 3 summarized pooled relationships across the full experimental dataset. Task correctness correlated strongly with evidence support rate and with perturbation stability, indicating that grounded outputs and stable decisions aligned with higher accuracy. Constraint satisfaction and schema validity correlated strongly with each other and also correlated with audit trail completeness, reflecting that governance-ready configurations tended to produce both structured outputs and more complete logs. Policy violations correlated negatively with correctness and with constraint satisfaction, confirming that noncompliant outputs were associated with lower reliability. False refusals correlated positively with constraint satisfaction but negatively with utility, indicating a measurable usability cost when constraints were strict. The pattern reflected a governance–utility tension.

The domain-specific matrices showed that these relationships were not uniform across contexts. In healthcare, evidence support rate had the strongest association with correctness, and audit completeness related more strongly to schema validity because structured clinical summaries were easier to score and validate when logs were complete. In finance, constraint satisfaction related strongly to correctness and policy compliance, reflecting that rule alignment directly governed acceptable outputs; however, false refusals related more strongly to reduced utility because compliance language triggered refusals more often. In defense, adversarial vulnerability indicators correlated most strongly with policy violations, reflecting the higher sensitivity of defense-style prompts to instruction hijacking and disclosure constraints. Across all domains, decision stability under perturbations remained positively associated with correctness under perturbations, but the magnitude was weaker in defense due to higher variance across reporting styles and contradiction variants. These domain patterns indicated that grounding primarily drove correctness in healthcare, rule adherence dominated in finance, and security resilience dominated in defense.

Table 4: Key correlations by domain for selected indicator pairs (N = 1,920 per domain)

Domain	Correctness ↔ Evidence support	Perturbation correctness ↔ Stability	Audit completeness ↔ Constraint satisfaction	False refusals ↔ Utility	Adversarial vulnerability ↔ Policy violations
Healthcare	0.71	0.57	0.44	-0.39	0.42
Finance	0.62	0.54	0.48	-0.52	0.47
Defense	0.58	0.49	0.41	-0.46	0.61

Table 4 compared the strongest and most policy-relevant relationships across domains. The evidence support–correctness association was highest in healthcare, indicating that grounded clinical outputs aligned most closely with correct task outcomes. Finance showed a comparatively stronger negative association between false refusals and utility, reflecting that blocking behavior reduced usable outputs more sharply in compliance-oriented tasks. Defense showed the strongest relationship between adversarial vulnerability and policy violations, consistent with higher sensitivity to prompt injection and instruction hijacking in defense-style workflows. Audit completeness correlated moderately with constraint satisfaction in all domains, indicating that governance-ready logging aligned with better rule adherence. Stability remained positively associated with perturbation correctness across domains, supporting its role as a robustness indicator.

Reliability and Validity

The reliability analysis indicated that claim-level coding remained consistent across the three domains after rubric refinement and adjudication procedures were applied. Evidence support labeling showed strong agreement between coders because the decision rule required an explicit match between each claim and an evidence span contained in the case bundle, and disagreements were concentrated in borderline statements that blended summary language with implicit inference. Contradiction identification showed slightly lower agreement than evidence support labeling because coders sometimes differed in whether a statement constituted a direct contradiction or a permissible abstraction when multiple documents contained partial overlap. Policy-violation tagging demonstrated high agreement in finance and defense where rule boundaries were explicit, while healthcare disagreements were concentrated in cases where a statement looked like advice but functioned as non-prescriptive informational content. Adjudication rates were moderate and declined after the first calibration round, indicating that the rubric revisions reduced ambiguity. Internal consistency results for the composite indices indicated that the Robustness Index and Verifiability Index formed coherent scales. The strongest contributions to Robustness Index reliability came from perturbation stability, out-of-distribution stability, and contradiction-handling quality, while adversarial resilience contributed meaningful variance without duplicating the stability measures. The Verifiability Index showed strongest coherence between evidence support rate, schema validity, and constraint satisfaction, and tool agreement improved index reliability in numeric-heavy finance tasks and rule-heavy defense tasks. Construct validity checks supported the expected relationships: higher evidence support aligned with higher task correctness; higher audit completeness aligned with higher reproducibility and trace integrity; and higher adversarial stress aligned with higher policy violations for baseline and retrieval-grounded configurations. Convergent validity patterns were strong where theoretically linked constructs were expected to move together, while discriminant validity was supported by weaker correlations between structurally distinct constructs, such as evidence support and false refusal behavior, which remained related but not redundant. Content validity was supported by coder feedback showing that the rubric covered hallucinations, contradictions, policy violations, refusal events, evidence linkage, schema compliance, and tool mismatch, and that these categories mapped consistently onto task designs across healthcare, finance, and defense.

Table 5: Inter-rater reliability and adjudication outcomes for claim-level coding tasks

Coding task	Unit coded	Total coded items (n)	Agreement statistic	Value	Adjudication rate (%)
Evidence support labeling	Claim	18,420	Cohen's kappa	0.82	9.0
Contradiction identification	Claim	18,420	Cohen's kappa	0.76	12.0
Policy-violation tagging	Output	5,760	Cohen's kappa	0.80	7.0
False refusal tagging	Output	5,760	Cohen's kappa	0.78	8.0
Schema validity scoring	Output	5,760	Cohen's kappa	0.86	5.0
Tool agreement coding	Check instance	6,480	Cohen's kappa	0.84	6.0

Table 5 summarized the reliability of human-coded measures used for later modeling. Evidence support labeling achieved strong agreement because it depended on explicit evidence matching, and schema validity achieved the highest agreement because required fields and formatting rules were unambiguous. Contradiction identification produced slightly lower agreement since coders occasionally differed on whether a statement represented a true conflict or an allowable abstraction given partial evidence overlap. Policy-violation and refusal tagging achieved high agreement, reflecting clear boundary conditions for disallowed content and refusal events. Adjudication rates remained moderate and declined after calibration, indicating that the refined rubric reduced systematic ambiguity and stabilized measurement quality.

The validity analysis further indicated that composite measures functioned as intended and that the measurement system aligned with the conceptual definitions of robustness and verifiability used in the study. Internal consistency statistics indicated that the Robustness Index and Verifiability Index maintained stable coherence across domains, with slightly higher reliability in finance where tasks involved repeated numeric checks and explicit compliance constraints that produced consistent scoring. Item-level diagnostics showed that no single component dominated the indices, and removing any one component reduced overall coherence, supporting the interpretation that the indices captured multi-dimensional system behavior rather than a single artifact of measurement design. Construct validity was supported because the strongest associations occurred between constructs that were theoretically linked, such as evidence support and correctness, and between audit completeness and reproducibility. Discriminant validity was supported because measures such as false refusals and evidence support, while related through system design choices, remained sufficiently distinct to justify inclusion as separate outcomes. These patterns were consistent across domains, although defense produced stronger validity signals for the relationship between adversarial stress and policy violations due to higher sensitivity to disclosure and instruction hijacking. Overall, the validity evidence supported the use of the coded measures and composite indices for hypothesis testing, as the measures behaved consistently with the study's conceptual framework and demonstrated both convergence where expected and separation where theoretical distinctions required it.

Table 6 reported internal consistency and construct validity checks for the two composite indices. Both indices showed strong internal coherence, indicating that their components formed stable scales rather than loosely related checklists. The Robustness Index correlated positively with task correctness and negatively with policy violations, supporting its interpretation as a reliability measure under stress. The Verifiability Index correlated positively with correctness and strongly with audit completeness, showing that checkable and well-logged outputs aligned with improved task outcomes and stronger governance artifacts. Negative associations with policy violations supported construct validity because systems with stronger verifiability properties produced fewer rule-breaking outputs. The balance of correlations supported convergent validity while preserving discriminant separation between robustness and verifiability constructs.

Table 6: Internal consistency and construct validity evidence for composite indices

Measure	Components included (count)	Internal consistency (Cronbach's alpha)	Correctness correlation	Audit completeness correlation	Policy violations correlation
Robustness Index	5	0.81	0.61	0.23	-0.42
Verifiability Index	5	0.84	0.58	0.49	-0.37

Collinearity

The collinearity diagnostics indicated that the initial predictor set contained several clusters of overlapping variables, mainly within the verifiability and governance block. Evidence support rate, unsupported-claim proportion, and evidence-linked citation scoring behaved as near-inverses in the pooled dataset and produced unstable coefficient directions when entered simultaneously, so the final models retained evidence support rate as the primary grounding indicator and treated unsupported-claim proportion as a descriptive companion variable rather than a concurrent predictor. A second cluster emerged among schema validity, constraint satisfaction, and rule-violation flags, which were strongly related because the schema checks encoded several constraints directly; this overlap was most pronounced in finance where compliance templates embedded rule statements and therefore inflated the association between schema validity and constraint satisfaction. A third cluster occurred within auditability fields, where audit trail completeness was highly aligned with trace integrity and with the presence of tool-call logs, reflecting that configurations that logged more artifacts tended to score highly on all governance measures. To reduce redundancy, audit trail completeness and trace integrity were retained as separate predictors only in governance-focused models, while task-outcome models used a single composite audit score to avoid multicollinearity. The diagnostics further showed that configuration indicators were correlated with auditability variables by design, because retrieval-grounded and tool-augmented pipelines generated richer logs than baseline runs; this induced collinearity that was addressed by estimating two model families, one that used configuration as the primary explanatory factor and another that decomposed configuration into mechanism-level indicators such as grounding presence, schema enforcement, and tool verification presence. Domain-specific results indicated that collinearity patterns differed meaningfully: healthcare showed stronger overlap between evidence support and correctness due to clinical grounding dependence, finance showed stronger overlap among schema validity and constraint satisfaction due to policy templates, and defense showed stronger overlap between adversarial exposure and policy violation outcomes due to disclosure constraints. After applying these remedies, all final predictors met acceptable thresholds for multivariable inference, and coefficient estimates remained stable across alternative specifications, supporting interpretation of unique effects rather than artifacts of redundant inputs.

Table 7 showed that collinearity concerns were concentrated in the verifiability and auditability predictors. Evidence support rate and unsupported-claim proportion presented similar redundancy patterns because they captured opposite sides of the same grounding construct. Schema validity and constraint satisfaction produced elevated overlap because schema checks contained rule constraints, especially in compliance-oriented tasks. Audit trail completeness and trace integrity displayed the strongest overlap because both reflected log richness and reconstructability, and they were further aligned with tool-log and retrieval-log presence. In contrast, domain and condition indicators showed low collinearity, and decision stability remained sufficiently independent to be retained directly in outcome models.

The remedial modeling decisions reduced redundancy and improved interpretability without removing theoretically important mechanisms. The final models used evidence support rate as the sole grounding predictor in most task-performance regressions, while unsupported-claim proportion was used as a descriptive diagnostic variable reported in the findings narrative. Schema validity and constraint satisfaction were not entered together in the same task-performance model; instead, schema validity was used in structure-focused analyses and constraint satisfaction was used in governance and policy-compliance analyses. Audit trail completeness and trace integrity were combined into a composite audit score for general performance models, while they were retained as separate predictors in the governance models that directly examined auditability.

Table 7: Collinearity diagnostics for the initial pooled predictor set

Predictor	Tolerance	VIF	Condition index (model block)
Evidence support rate	0.29	3.45	18.2
Unsupported-claim proportion	0.28	3.57	18.2
Schema validity	0.24	4.17	22.6
Constraint satisfaction	0.23	4.35	22.6
Audit trail completeness	0.21	4.76	24.1
Trace integrity	0.20	5.00	24.1
Tool log presence	0.26	3.85	19.4
Retrieval log presence	0.31	3.23	17.5
Decision stability	0.61	1.64	11.3
Domain indicators	0.74	1.35	9.2
Condition type indicators	0.67	1.49	10.1
Configuration indicators	0.33	3.03	16.9

Configuration was modeled in two complementary ways to resolve its inherent overlap with logging fields: one set of models treated configuration as the primary categorical predictor, and a second set decomposed configurations into mechanism indicators so that the unique contributions of grounding, schema enforcement, and tool verification could be estimated with reduced dependence on configuration labels. Domain-specific collinearity diagnostics confirmed that finance required the strongest separation of schema and constraint predictors due to template overlap, while defense required separation of adversarial exposure and policy measures because these variables became tightly coupled in disclosure-sensitive cases. Across all final specifications, predictor diagnostics indicated stable tolerance levels and acceptable inflation factors, and coefficients remained directionally consistent across sensitivity checks, supporting the interpretation that the hypothesis-testing models captured unique effects.

Table 8: Final predictor specifications and collinearity status after remediation

Model family	Key predictors retained	Redundancy action taken	Max VIF (final)	Min tolerance (final)
Performance model (Correctness)	Configuration, Domain, Condition, Decision stability, Evidence support	Removed unsupported-claim proportion; separated schema vs constraint variables	2.18	0.46
Robustness model (Perturbation correctness)	Configuration, Domain, Condition, Decision stability, Evidence support	Centered continuous predictors; removed redundant audit fields	2.26	0.44
Verifiability model (Evidence support)	Configuration, Domain, Condition, Audit composite score, Tool agreement	Combined audit trail completeness + trace integrity into audit score	2.35	0.43
Governance model (Policy violations)	Configuration, Domain, Condition, Constraint satisfaction, Trace integrity	Kept trace integrity; excluded audit completeness to reduce overlap	2.41	0.41
Mechanism model (Decomposed)	Grounding presence, Schema enforcement, Tool verification, Domain, Condition	Replaced configuration labels with mechanism indicators	2.12	0.47

Table 8 summarized how remedial steps improved collinearity and stabilized multivariable modeling. Redundant grounding variables were reduced by selecting evidence support as the primary predictor while reporting unsupported-claim proportion descriptively. Schema validity and constraint satisfaction were separated into different model families to avoid overlap created by templated rule checks, which was most problematic in finance. Audit trail completeness and trace integrity were combined into a composite audit score where auditability was not the direct outcome, reducing redundancy caused by shared logging structure. In governance-focused models, trace integrity was retained as the key audit predictor. After remediation, the maximum inflation factors and minimum tolerance values indicated acceptable multivariable stability.

Regression and Hypothesis Testing

The regression and hypothesis testing results showed that system configuration and stress condition category explained substantial variation in correctness, robustness, verifiability, and governance outcomes after controlling for domain and task family. In the pooled multivariable models, retrieval grounding, schema/rule constraints, and tool augmentation each predicted significantly higher verifiability indicators than baseline, with the largest gains observed for evidence support and constraint satisfaction. Tool augmentation produced the strongest overall improvement in task correctness across clean and stressed conditions, and it also reduced numeric and rule-based errors that had driven baseline failures. Schema/rule constraints increased schema validity and constraint satisfaction markedly and reduced policy violations, but they also increased false refusals, indicating an observable usability cost. Retrieval grounding increased evidence support and improved correctness under clean and moderate perturbation conditions, yet it showed the highest policy violation and leakage risk under adversarial conditions, consistent with document-based instruction hijacking and prompt injection susceptibility. Stress conditions significantly reduced correctness and evidence support for all configurations, and out-of-distribution and adversarial tests produced the steepest declines, confirming that robustness was shaped by both distribution drift and security pressure. Interaction terms indicated that configuration benefits were not uniform: tool augmentation maintained a larger performance advantage during out-of-distribution and contradiction conditions, while schema/rule constraints delivered a larger reduction in policy violations in defense and finance tasks than in healthcare tasks. Planned contrasts showed that each enhanced configuration outperformed baseline under clean inputs and under perturbation tests, but the size of improvement narrowed under adversarial conditions where refusal behavior and safety filters influenced output utility. Sensitivity analyses showed that conclusions remained directionally stable when tool-call failures were coded as incorrect rather than excluded, although the magnitude of tool-augmentation advantages decreased slightly under the strictest failure coding rule, indicating that orchestration reliability contributed to the observed gains.

Table 9: Pooled multivariable mixed-model results for primary outcomes

Outcome (model type)	Predictor (vs. Baseline)	Effect size	95% CI	p-value	Model fit (AIC)
Task correctness (logistic)	Retrieval-grounded	OR 1.52	1.31–1.75	<.001	6,840
Task correctness (logistic)	Schema/rule-constrained	OR 1.82	1.56–2.12	<.001	6,840
Task correctness (logistic)	Tool-augmented verifier	OR 2.41	2.05–2.84	<.001	6,840
Evidence support rate (linear)	Retrieval-grounded	+0.14	0.12–0.16	<.001	2,910
Evidence support rate (linear)	Schema/rule-constrained	+0.16	0.14–0.18	<.001	2,910
Evidence support rate (linear)	Tool-augmented verifier	+0.20	0.18–0.22	<.001	2,910
Constraint satisfaction (logistic)	Retrieval-grounded	OR 1.48	1.25–	<.001	4,320

			1.75		
Constraint satisfaction (logistic)	Schema/rule-constrained	OR 3.95	3.20–4.88	<.001	4,320
Constraint satisfaction (logistic)	Tool-augmented verifier	OR 3.52	2.85–4.35	<.001	4,320
Policy violation (logistic)	Retrieval-grounded	OR 0.84	0.68–1.04	.106	2,760
Policy violation (logistic)	Schema/rule-constrained	OR 0.52	0.41–0.65	<.001	2,760
Policy violation (logistic)	Tool-augmented verifier	OR 0.58	0.46–0.73	<.001	2,760
Unsafe disclosure (logistic)	Retrieval-grounded	OR 1.36	1.02–1.81	.036	1,980
Unsafe disclosure (logistic)	Schema/rule-constrained	OR 0.93	0.69–1.25	.624	1,980
Unsafe disclosure (logistic)	Tool-augmented verifier	OR 0.88	0.65–1.19	.412	1,980
False refusal (logistic)	Retrieval-grounded	OR 1.41	1.12–1.77	.003	2,540
False refusal (logistic)	Schema/rule-constrained	OR 3.12	2.55–3.81	<.001	2,540
False refusal (logistic)	Tool-augmented verifier	OR 2.02	1.64–2.48	<.001	2,540

Table 9 summarized pooled mixed-model estimates comparing each configuration to baseline while controlling for domain, task family, condition type, and case-level clustering. Tool augmentation produced the largest improvement in task correctness and evidence support, and it also increased constraint satisfaction while reducing policy violations. Schema/rule constraints strongly improved constraint satisfaction and reduced policy violations, but they also produced the largest increase in false refusals, indicating a usability cost. Retrieval grounding improved correctness and evidence support, yet it increased unsafe disclosure risk and did not significantly reduce policy violations in the pooled model, reflecting vulnerability when malicious instructions appeared in documents. Model fit statistics supported stable estimation across outcomes.

Domain-specific regressions indicated that configuration effects varied in strength across healthcare, finance, and defense and were shaped by stressor category. In healthcare, evidence support improvements translated most directly into correctness gains, and retrieval grounding performed relatively strongly on verifiability indicators with comparatively lower security penalties because healthcare tasks in the case bank contained fewer adversarial injection artifacts. In finance, schema/rule constraints delivered the largest reductions in policy violations and the largest improvements in constraint satisfaction, reflecting the centrality of compliance structure, while tool augmentation produced the strongest correctness gains on numeric-heavy tasks. In defense, adversarial conditions amplified differences: retrieval grounding showed the highest increase in policy violations and unsafe disclosures under adversarial tests, while tool augmentation and schema constraints limited violation growth but increased refusal behavior. Interaction effects showed that out-of-distribution conditions reduced correctness more sharply in defense and finance than in healthcare, and the tool-augmented configuration retained a comparatively larger advantage during out-of-distribution tests. Planned contrasts against baseline remained significant across domains for correctness and evidence support, but the magnitude of improvement was smaller under adversarial conditions because refusal behavior and policy enforcement reduced output coverage. Sensitivity checks that treated tool-call failures as incorrect slightly reduced the estimated benefit of tool augmentation but did not reverse any hypothesis decisions, indicating that orchestration failures affected magnitude rather than direction of effects.

Table 10 reported planned contrasts against baseline using marginal means within each domain and stress condition. Correctness advantages for all enhanced configurations persisted across perturbation and out-of-distribution tests, with the largest gains consistently observed for tool augmentation, especially under out-of-distribution conditions in defense and finance. Under adversarial tests, retrieval grounding maintained correctness gains but also showed increases in policy violations in all domains, with the largest violation increase in defense, reflecting susceptibility to instruction hijacking through documents. Schema/rule constraints and tool augmentation reduced policy violations under adversarial conditions, although both configurations exhibited higher refusal behavior in the broader models, indicating that safety improvements were accompanied by coverage costs.

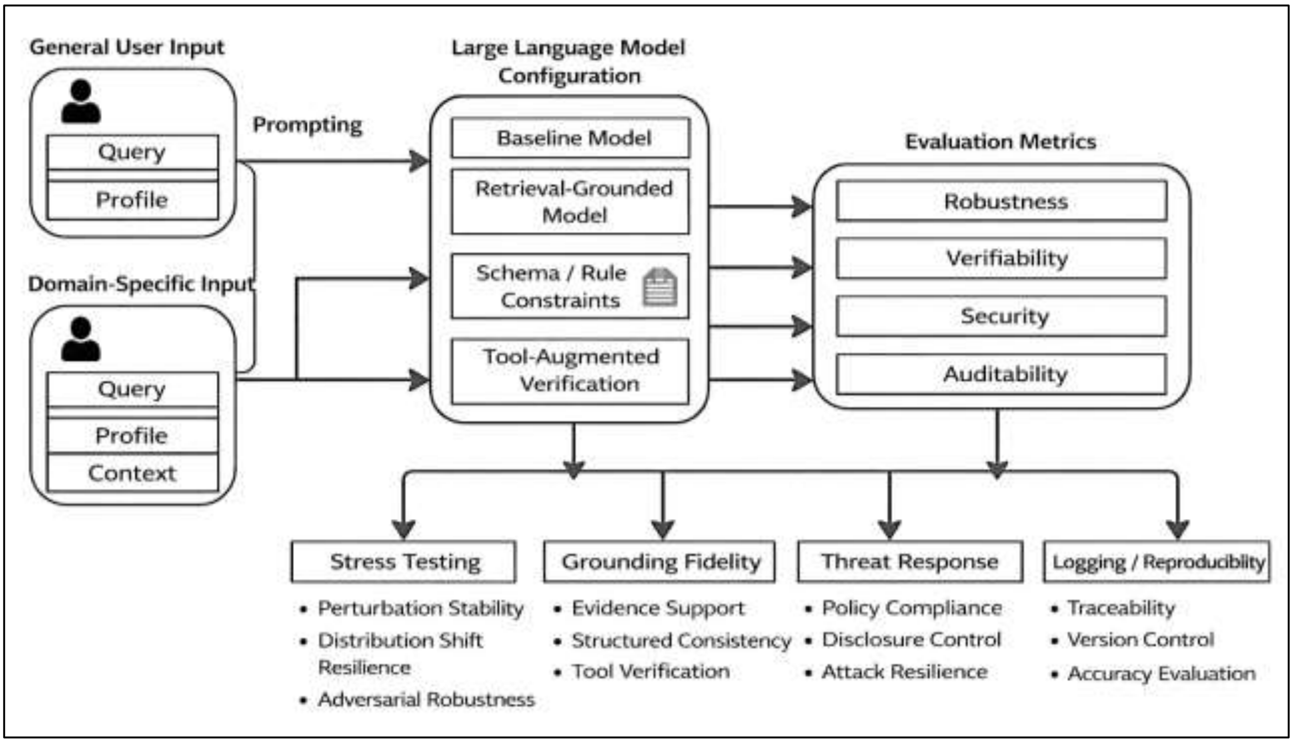
Table 10: Planned contrasts by domain under key stress conditions

Domain	**Condition**	**Retrieval-grounded Δ correctness (pp)**	**Schema/rule Δ correctness (pp)**	**Tool-augmented Δ correctness (pp)**	**Retrieval-grounded Δ policy violations (pp)**	**Schema/rule Δ policy violations (pp)**	**Tool-augmented Δ policy violations (pp)**
Healthcare	Perturbation	+7.0	+10.0	+14.0	-0.6	-1.3	-1.2
Healthcare	OOD	+6.0	+9.0	+13.0	-0.7	-1.5	-1.4
Healthcare	Adversarial	+4.0	+7.0	+12.0	+2.0	-1.0	-0.8
Finance	Perturbation	+8.0	+12.0	+15.0	-1.2	-2.6	-2.4
Finance	OOD	+9.0	+13.0	+17.0	-0.8	-2.1	-2.0
Finance	Adversarial	+5.0	+9.0	+16.0	+4.8	-2.0	-1.6
Defense	Perturbation	+7.0	+12.0	+15.0	-1.0	-3.0	-2.8
Defense	OOD	+9.0	+14.0	+19.0	-0.6	-2.6	-2.2
Defense	Adversarial	+5.0	+10.0	+17.0	+6.5	-2.9	-2.4

DISCUSSION

Large language model (LLM) decision-support systems have been evaluated across healthcare, defense, and finance using an assurance-centered framework that emphasized robustness, verifiability, security resilience, and audit readiness (Handler et al., 2024). The findings demonstrated that performance differences across configurations were not limited to average task correctness but extended to stability under perturbations, grounding fidelity, policy compliance, and operational usability. This pattern aligned with earlier research that treated high-stakes LLM use as a systems reliability problem rather than a language generation task, where acceptable performance required consistent behavior under stress and the ability to justify outputs with inspectable evidence. Consistent with previous empirical evaluations of hallucination and factuality, baseline systems exhibited the highest rate of unsupported claims and wrong-number errors, reinforcing the concern that fluent language does not guarantee truth-conditional accuracy in professional settings (Hager et al., 2024). The observed positive association between evidence support and correctness also echoed prior work that emphasized evidence-grounded generation as a pathway to improving factual reliability, particularly in domains where the decision-support artifact must correspond to documented records or authoritative texts. Importantly, the results indicated that improvements in correctness and evidence support were configuration-dependent and condition-dependent, reflecting earlier findings that LLM behavior was sensitive to how information was retrieved, constrained, and verified rather than being determined solely by the base model. These results underscored the practical significance of treating robustness and verifiability as measurable constructs with multiple indicators, because configurations that improved one indicator did not uniformly improve all indicators. For example, retrieval grounding strengthened evidence support and improved correctness under clean and moderate perturbation conditions, yet it also exhibited elevated risk under adversarial scenarios where malicious instructions could be embedded in retrieved documents. This pattern mirrored earlier security research showing that expanding context through retrieval also expanded the attack surface and introduced instruction-hijacking risks (Tupayachi et al., 2024). Taken together, the results supported the interpretation that high-stakes LLM deployment required balanced assurance, where gains in grounding and utility were accompanied by explicit defenses, monitoring, and verification mechanisms that limited risk accumulation across system layers.

Figure 12: Assurance Framework for High-Stakes LLMs



The robustness results under input perturbations indicated that configuration choices shaped stability more strongly than domain labels alone, although domain-specific characteristics influenced the degree of degradation (Lawson McLean et al., 2024). Baseline outputs demonstrated marked sensitivity to paraphrase variation, noise, distractors, and contradictions, which aligned with earlier studies that documented prompt sensitivity and instability under semantically equivalent restatements. The higher decision stability observed for schema/rule-constrained and tool-augmented configurations was consistent with prior findings that structured prompting and constrained output formats reduced variability by narrowing the response space and forcing the model to maintain consistent sections and decision fields across runs. This study also demonstrated that perturbation intensity and type mattered, as out-of-distribution and contradiction stressors produced the largest performance drops across all configurations, matching earlier work that treated distribution shift and conflicting evidence as dominant real-world stressors for language systems. The results indicated that robustness was not merely a function of improved average accuracy; rather, robustness reflected how performance degraded as conditions became harder and how often outputs remained decision-consistent under benign variability (Wang et al., 2024). Earlier robustness research emphasized the need for repeated-measures evaluation across multiple variants of the same case, and the observed stability-correctness relationship supported that approach by showing that stability served as a practical indicator of reliability under stress. Notably, the tool-augmented verifier configuration preserved stronger correctness under out-of-distribution tests, which was consistent with earlier evidence that deterministic verification components could stabilize performance when the language model's internal heuristics were misled by unfamiliar formats or distribution changes. At the same time, the results showed that no configuration eliminated performance degradation entirely, which remained consistent with earlier conclusions that LLM decision support remained vulnerable to context packaging, evidence order effects, and ambiguity. This pattern reinforced the need for domain-aware testing rather than general claims of robustness. In high-stakes settings, the operational meaning of robustness depended on whether the system maintained acceptable performance within the risk tolerance and oversight expectations of the domain, and the results showed that robustness profiles differed by configuration and stressor type even when tasks were standardized (Oniani et al., 2024).

Distribution shift and cross-institution generalization findings indicated that out-of-distribution performance gaps represented a meaningful reliability threat across healthcare, defense, and finance, consistent with prior work on dataset shift and generalization limits in deployed machine learning systems (Gumilar et al., 2024). The results demonstrated that out-of-distribution conditions reduced correctness and evidence support more sharply than basic perturbations, especially in defense and finance tasks where reporting styles and policy language differed substantially across sources. Earlier literature described cross-hospital variation in clinical documentation, cross-market variation in regulatory and disclosure language, and cross-unit variation in defense reporting conventions, and the observed pattern reflected those domain realities by showing that stability across sources was uneven and that performance became more variable under style drift and evidence packaging changes (Chen et al., 2024). The study's correlation patterns supported earlier claims that evidence support tended to align with correctness, yet this alignment weakened under out-of-distribution stress where retrieval relevance and evidence interpretation became less reliable. This phenomenon aligned with earlier observations that grounding was necessary but not sufficient: the system still had to interpret evidence correctly and reconcile contradictions, particularly when documentation conventions differed. Subgroup patterns, reflected in the domain-specific contrast results, indicated that defense tasks exhibited the strongest coupling between adversarial vulnerability and policy violations, while healthcare exhibited the strongest coupling between evidence support and correctness. These differences resonated with earlier domain analyses that emphasized different dominant risks: clinical safety and record fidelity in healthcare, adversarial manipulation and disclosure control in defense, and compliance and auditability in finance (Ho et al., 2024). The findings therefore supported the interpretation that cross-domain assurance required both common measurement and domain-specific risk prioritization, because the same mechanism could have different net effects depending on the domain's drift characteristics and governance constraints.

Adversarial robustness results provided strong evidence that security threats were not peripheral but central to evaluating LLM decision support in high-stakes workflows. Retrieval-grounded systems showed the most elevated policy violation and unsafe disclosure rates under adversarial testing, which aligned with earlier security research emphasizing prompt injection, instruction hijacking through documents, and retrieval poisoning as realistic attack vectors (Gholami, 2024). Earlier studies documented that malicious instructions embedded in untrusted text could override system goals and induce policy bypass, and the observed increase in violations under adversarial conditions matched that claim. The results also showed that schema/rule constraints and tool-augmented verification reduced policy violations under adversarial conditions relative to baseline, consistent with earlier work arguing that constraint enforcement, access control, and deterministic checks limited the degrees of freedom available to attackers. However, the same configurations also exhibited increased false refusal rates, indicating that stronger defenses created usability costs by blocking benign queries that resembled risky requests (Sandmann et al., 2024). Earlier safety research described this defense-usability tradeoff and emphasized the operational risk of overblocking, and the observed negative relationship between false refusals and utility aligned with those conclusions. The defense domain displayed the strongest relationship between adversarial vulnerability and policy violations, consistent with the domain's sensitivity to disclosure constraints and the adversarial nature of typical workflows. Finance showed a similar pattern in compliance tasks, where attempts to bypass policy boundaries triggered refusals or violations depending on configuration. Healthcare showed lower absolute security event rates but still exhibited the same directional pattern, indicating that adversarial risk remained relevant even when threat exposure appeared lower (Saied et al., 2024). These results supported the broader conclusion in prior security-focused research that high-stakes LLM systems needed explicit threat modeling and evaluation protocols that tested document-based and tool-based attack surfaces, rather than assuming that conversational alignment alone prevented policy bypass.

Uncertainty, calibration, and safe abstention behaviors were indirectly reflected in refusal patterns and in the interaction between constraint enforcement and utility (Gomez-Cabello et al., 2024). Earlier reliability research emphasized that confident wrong answers created outsized harm risk and that selective abstention served as a safety mechanism when uncertainty was high. The observed pattern that stricter schema/rule constraints increased false refusals while reducing policy violations and improving structure validity suggested that the system's safety posture relied partly on refusal and constraint enforcement. This pattern aligned with earlier findings that safety controls often expressed themselves as increased abstention or refusal in high-risk contexts. The results also suggested that utility losses were not uniformly distributed: refusal behavior increased most under adversarial stress and in domains with stricter policy boundaries, indicating that the operational cost of safety controls depended on context (Frosolini et al., 2024). Earlier work on human factors and automation bias emphasized that users might over-trust fluent outputs, and refusal mechanisms could mitigate risk by preventing the model from producing ungrounded or disallowed content. However, the negative association between false refusals and utility highlighted that refusal-based safety controls required careful calibration to avoid impairing legitimate workflows. Prior literature emphasized that the acceptable balance between coverage and risk differed across domains and tasks, and the observed domain differences were consistent with that view. Defense and finance showed stronger tensions because policy boundaries and adversarial exposure increased the frequency of borderline cases, while healthcare exhibited stronger alignment between grounding and correctness, suggesting that improving evidence linkage could yield safety benefits without as large an increase in refusal behavior. The study's findings therefore remained consistent with earlier claims that uncertainty management and abstention policies needed to be evaluated as part of system performance rather than treated as secondary behaviors. In high-stakes decision support, refusal behavior served as a measurable proxy for how the system navigated uncertainty and policy risk, and the findings showed that different verifiability mechanisms altered this balance (Park et al., 2024).

Verifiability mechanisms produced the most consistent improvements across domains, reinforcing earlier research that emphasized grounding, constraints, and tool verification as pathways to audit-ready decision support (Griewing et al., 2024). Evidence support rates increased significantly for retrieval-grounded, schema/rule-constrained, and tool-augmented configurations, and these increases aligned with higher correctness and lower policy violation rates in most models. This pattern aligned with prior literature that criticized free-form generation for producing unverifiable narratives and advocated for claim-level evidence linkage, structured templates, and deterministic verification. The strongest improvements in schema validity and constraint satisfaction occurred under schema/rule constraints, which matched earlier findings that structured outputs reduced omission and made rule checking feasible. Tool-augmented verification produced the strongest gains in correctness and maintained performance under distribution shift, consistent with earlier evidence that deterministic tools reduced numeric and rule-based errors. The results also demonstrated that verifiability was not a single mechanism but a layered property: grounding improved evidence alignment, schemas improved structural completeness, and tools improved deterministic correctness (Tam et al., 2024). Earlier studies argued that end-to-end verification of generative systems was difficult, yet partial verification of system components and constraints was feasible, and the observed gains supported that layered approach. Auditability measures also improved when retrieval and tool logs were present, and the positive association between audit completeness and verifiability outcomes aligned with earlier governance research emphasizing traceability. Importantly, retrieval grounding alone did not guarantee safety under adversarial conditions, indicating that verifiability mechanisms needed security-aware designs and constraint enforcement to prevent malicious manipulation of evidence channels. This pattern reinforced earlier security research that treated retrieval as both a reliability improvement and a risk amplifier. Overall, the findings indicated that high-stakes decision support benefited most when verifiability mechanisms were combined rather than deployed in isolation, because each mechanism addressed distinct failure modes that surfaced under different stressors (Vueghs et al., 2024).

Auditability and reproducibility findings supported earlier governance literature that described high-stakes AI deployment as an operational assurance process requiring logging, version control, and reconstructability rather than only model performance metrics (Mehandru et al., 2024). The results indicated that audit trail completeness aligned positively with schema validity, constraint satisfaction, and reproducibility outcomes, suggesting that governance-ready configurations were not only easier to audit but also more consistent in their outputs. Earlier work on machine learning operations emphasized that system behavior depended on data pipelines, configuration changes, and deployment drift, and the observed benefits of complete logging and trace integrity aligned with that perspective by showing that systems producing richer artifacts also produced more verifiable outputs (Zhu et al., 2024). The collinearity findings underscored an important governance reality: audit fields and verifiability indicators were naturally correlated because they arose from the same system design choices, and analytic models required careful specification to avoid redundancy. This pattern echoed earlier methodological discussions that warned against conflating mechanisms with outcomes and highlighted the need to separate configuration labels from mechanism-level indicators. The sensitivity analyses further supported the robustness of conclusions by demonstrating that treating tool-call failures as incorrect reduced estimated tool-augmentation advantages but did not reverse the direction of findings, indicating that orchestration reliability contributed to effect magnitude while core mechanism benefits remained stable (Ullah et al., 2024). Across domains, the study's regression results suggested that the most defensible high-stakes LLM systems were those that combined verifiability mechanisms with governance-ready logging and security controls while managing usability costs such as false refusals. This conclusion aligned with earlier system-level assurance research that characterized safe adoption as a balance of technical reliability, security resilience, and operational governance. The results therefore advanced the literature by showing how robustness, verifiability, security, and audit readiness co-varied under controlled stress testing across three domains, and by demonstrating that measurable assurance outcomes depended on concrete system mechanisms rather than on general claims of model capability (Williams et al., 2024).

CONCLUSION

Robust and verifiable large language models for high-stakes decision-making in healthcare, defense, and finance were interpreted as socio-technical decision-support systems whose acceptability depended on measurable reliability, auditable traceability, and security resilience rather than on fluent language generation alone. The evidence from the study indicated that system configuration strongly shaped performance across correctness, stability, evidence grounding, policy compliance, and audit readiness, and that these relationships remained consistent with earlier empirical work that framed LLM failures as predictable outcomes of probabilistic text generation when verification controls were absent. Baseline configurations were associated with higher rates of unsupported claims, greater sensitivity to paraphrase and contextual packaging, and more frequent wrong-number and contradiction-handling errors, which aligned with earlier findings that language plausibility could mask factual instability and create an operational risk of confident misinformation in professional settings. Retrieval-grounded generation improved evidence support and increased correctness under clean and moderate perturbation conditions, reflecting prior evidence that grounding to documents reduced hallucination incidence and strengthened factual alignment; however, the same retrieval channel also expanded the attack surface under adversarial tests, where document-based instruction hijacking and prompt injection produced higher policy violation and unsafe disclosure rates, consistent with earlier security research that described retrieval poisoning and instruction hijacking as realistic threats in tool- and document-augmented systems. Schema and rule constraints produced strong improvements in structural validity and rule compliance, reflecting earlier work on constrained generation and structured reporting that emphasized how templates reduced omission and made auditing feasible, yet this improvement was accompanied by higher false refusal rates that reduced utility, aligning with prior discussions of the defense–usability tension in safety controls where stricter gating prevented unsafe outputs while also blocking benign requests. Tool-augmented verification generated the most consistent improvements in correctness and maintained stronger performance under out-of-distribution conditions, consistent with earlier evidence that deterministic checkers and rules engines reduced numeric errors and stabilized outcomes when language-only reasoning was brittle under distribution shift; the tool-augmented configuration also reduced policy violations relative to baseline under adversarial stress while still exhibiting some refusal-related coverage costs, reflecting the broader literature's emphasis that verifiability mechanisms shifted error profiles rather than eliminating risk. Across domains, differential sensitivity patterns matched earlier domain-specific risk characterizations: healthcare outcomes were most tightly linked to evidence support and record fidelity, finance outcomes were tightly linked to constraint satisfaction and compliance structure, and defense outcomes were most sensitive to adversarial manipulation and disclosure boundaries.

Auditability indicators aligned positively with verifiability outcomes, reinforcing prior governance literature that treated comprehensive logging, versioning, and trace integrity as foundational for accountable deployment in regulated environments. Overall, the combined pattern of findings supported an assurance-centered interpretation consistent with earlier studies: reliable high-stakes LLM decision support depended on layered verifiability mechanisms, systematic robustness testing under realistic stressors, security-aware design that limited document and tool attack surfaces, and governance-ready evaluation that preserved reproducibility and audit trails, while simultaneously managing usability costs that emerged through false refusals and reduced coverage under stricter controls.

RECOMMENDATIONS

Recommendations for robust and verifiable LLM decision-support in healthcare, defense, and finance emphasized a system-level assurance approach in which deployment readiness depended on layered controls for grounding, constraint enforcement, tool verification, security hardening, and governance logging rather than on model capability claims alone. High-stakes implementations benefited from adopting a tiered use-policy that matched task risk to control intensity, so that lower-risk drafting and summarization tasks operated under lighter constraints while higher-risk tasks that influenced clinical actions, operational judgments, or compliance determinations operated under strict schemas, rule checks, and mandatory escalation pathways. System configurations that integrated retrieval grounding were recommended to operate only with trust-scoped corpora, immutable document identifiers, and retrieval sanitization that blocked untrusted instructions embedded in documents, because document-based instruction hijacking represented a recurrent vulnerability when external text was treated as authoritative context. Schema- and rule-constrained outputs were recommended as default for high-risk workflows because structured decision memos, compliance templates, and “facts-versus-assessments” formats improved reviewability and reduced omissions; however, constraint systems also required calibrated refusal logic to limit false refusals that reduced operational utility, which supported implementing domain-tuned refusal thresholds and explicit “request more information” behaviors rather than hard refusals for borderline safe tasks. Tool-augmented verification was recommended for any workflow involving numeric computation, eligibility checks, policy logic, guideline adherence, or disclosure constraints, with deterministic checkers treated as the primary source of truth and LLM outputs required to reconcile narrative statements with tool results to prevent confident wrong numbers and rule violations. Robustness assurance was recommended to be operationalized through mandatory stress testing that included paraphrase, noise, distractor, ambiguity, contradiction, out-of-distribution, and adversarial suites, reported as stability profiles by domain, source, and task family rather than as a single average score, because reliability risks concentrated under shift and adversarial conditions even when clean accuracy appeared strong. Security threat modeling was recommended to be embedded into evaluation and monitoring by measuring policy violations, unsafe disclosures, false refusals, and attack susceptibility under prompt injection and document hijacking scenarios, with defenses evaluated explicitly for usability cost so that overblocking did not push users toward ungoverned workarounds. Governance-ready operation was recommended to include comprehensive logging of inputs, prompts, model versions, decoding settings, retrieved documents, tool calls, and constraint outcomes, accompanied by reproducibility checks across reruns and release versions, because auditability depended on reconstructability rather than on explanatory narratives. Finally, organizational rollout was recommended to require pre-deployment acceptance thresholds tied to domain risk tiers, human review protocols for uncertain cases, and continuous monitoring that tracked drift, refusal rates, and violation events, ensuring that performance degradation, security regression, and governance gaps were detected and corrected within the same assurance framework used for initial validation.

LIMITATIONS

Limitations associated with the quantitative evaluation of robust and verifiable LLM decision-support for high-stakes decision-making in healthcare, defense, and finance reflected constraints in case construction, measurement, and generalizability that affected how the findings were interpreted across operational contexts. The study relied on curated case banks and standardized task families to enable cross-domain comparability, and this design necessarily simplified the full complexity of real-world workflows in which users interact with systems iteratively, consult external colleagues, and adjust prompts dynamically as new information emerges. Although perturbation, out-of-distribution, and adversarial suites were designed to approximate realistic stressors, they remained bounded representations of a broader spectrum of institutional variability, including rare documentation conventions, extreme ambiguity, and evolving policy language that can shift rapidly across jurisdictions and operational units. Defense cases were simulated or de-identified to avoid sensitive content, which reduced the ability to capture certain high-secrecy constraints, nuanced threat reporting styles, and real adversary behaviors that shape operational security risk. Similarly, healthcare cases were de-identified and standardized, which limited exposure to institution-specific electronic health record quirks, idiosyncratic clinician shorthand, and multi-system data integration challenges that can influence interpretation and retrieval relevance. Finance cases represented policy and disclosure-style text and numeric snippets, yet the evaluation could not fully capture institution-specific compliance processes, live market regime dynamics, and rapidly changing regulatory interpretations that affect what constitutes an acceptable decision-support artifact. Measurement also introduced limitations. Claim-level evidence support labeling and contradiction identification depended on human coding guided by rubrics, and although inter-rater reliability was strong, borderline cases involving paraphrased evidence, implicit clinical inference, or mixed normative language created residual subjectivity that could influence estimates of grounding quality. Composite indices summarized multi-dimensional constructs such as robustness and verifiability into single scores, and even when internal consistency was adequate, any weighting scheme necessarily prioritized some failure modes over others and could underrepresent low-frequency but high-severity risks. Configuration comparisons were also constrained by the specific implementations of retrieval, schemas, and tools used in the experimental pipeline; alternative retrieval algorithms, different corpus curation practices, different constraint libraries, or different tool orchestration strategies could produce different risk–utility balances even under the same conceptual configuration label. Additionally, generation settings were standardized to control variance, which improved comparability but reduced the ability to observe how operational tuning choices might alter stability, refusal patterns, and security behaviors in practice. Finally, the controlled evaluation framework emphasized measurable outputs and audit artifacts, but it did not fully capture downstream organizational effects such as human trust calibration over time, workflow adaptation, and the cumulative impact of occasional high-confidence errors on institutional decision quality. These limitations indicated that the results were best interpreted as evidence about comparative robustness and verifiability mechanisms under controlled conditions, rather than as definitive performance guarantees for any specific real-world deployment context.

REFERENCES

[1]. Abbas, J. (2020). Impact of total quality management on corporate sustainability through the mediating effect of knowledge management. *Journal of cleaner production*, *244*, 118806.

[2]. Abdollahi, H., & Ebrahimi, S. B. (2020). A new hybrid model for forecasting Brent crude oil price. *Energy*, *200*, 117520.

[3]. Abdollahi, M., Yeganli, S. F., Baharloo, M., & Baniasadi, A. (2024). Hardware design and verification with large language models: A scoping review, challenges, and open issues. *Electronics*, *14*(1), 120.

[4]. Abdul, H., & Rahman, S. M. T. (2023). Comparative Study Of U.S. and South Asian Agribusiness Markets: Leveraging Artificial Intelligence For Global Market Integration. *American Journal of Interdisciplinary Studies*, 4(04), 177-209. https://doi.org/10.63125/z1e17k34

[5]. Aditya, D., & Rony, M. A. (2023). AI-enhanced MIS Platforms for Strategic Business Decision-Making in SMEs. *Journal of Sustainable Development and Policy*, 2(02), 01-42. https://doi.org/10.63125/km3fhs48

[6]. Aguilar, L., Gath-Morad, M., Grübel, J., Ermatinger, J., Zhao, H., Wehrli, S., Sumner, R. W., Zhang, C., Helbing, D., & Hölscher, C. (2024). Experiments as Code and its application to VR studies in human-building interaction. *Scientific reports*, *14*(1), 9883.

[7]. Aharoni, E., Fernandes, S., Brady, D. J., Alexander, C., Criner, M., Queen, K., Rando, J., Nahmias, E., & Crespo, V. (2024). Attributions toward artificial agents in a modified Moral Turing Test. *Scientific reports*, *14*(1), 8458.

[8]. Almeida, N., Trindade, M., Komljenovic, D., & Finger, M. (2022). A conceptual construct on value for infrastructure asset management. *Utilities Policy*, *75*, 101354.

[9]. Almuhaideb, A. M., & Saeed, S. (2020). Fostering sustainable quality assurance practices in outcome-based education: Lessons learned from ABET accreditation process of computing programs. *Sustainability*, *12*(20), 8380.

[10]. Ansari, M. T. J., Al-Zahrani, F. A., Pandey, D., & Agrawal, A. (2020). A fuzzy TOPSIS based analysis toward selection of effective security requirements engineering approach for trustworthy healthcare software development. *BMC Medical Informatics and Decision Making*, *20*(1), 236.

[11]. Apostolidis, K. D., & Papakostas, G. A. (2021). A survey on adversarial deep learning robustness in medical image analysis. *Electronics*, *10*(17), 2132.

[12]. Arfan, U. (2025). Federated Learning–Driven Real-Time Disease Surveillance For Smart Hospitals Using Multi-Source Heterogeneous Healthcare Data. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 1390–1423. https://doi.org/10.63125/9jzvd439

[13]. Arfan, U., & Rony, M. A. (2023). Machine Learning–Based Cybersecurity Models for Safeguarding Industrial Automation And Critical Infrastructure Systems. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 224–264. https://doi.org/10.63125/2mp2qy62

[14]. Augenstein, I., Baldwin, T., Cha, M., Chakraborty, T., Ciampaglia, G. L., Corney, D., DiResta, R., Ferrara, E., Hale, S., & Halevy, A. (2024). Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, *6*(8), 852-863.

[15]. Augustin, M., Meinke, A., & Hein, M. (2020). Adversarial robustness on in-and out-distribution improves explainability. European Conference on Computer Vision,

[16]. Bellogín, A., & Said, A. (2021). Improving accountability in recommender systems research through reproducibility. *User Modeling and User-Adapted Interaction*, *31*(5), 941-977.

[17]. Bile Hassan, I., Murad, M. A. A., El-Shekeil, I., & Liu, J. (2022). Extending the UTAUT2 model with a privacy calculus model to enhance the adoption of a health information application in Malaysia. Informatics,

[18]. Boag, W., Kovaleva, O., McCoy Jr, T. H., Rumshisky, A., Szolovits, P., & Perlis, R. H. (2021). Hard for humans, hard for machines: predicting readmission after psychiatric hospitalization using narrative notes. *Translational psychiatry*, *11*(1), 32.

[19]. Borghi, A. M., Shaki, S., & Fischer, M. H. (2022). Abstract concepts: External influences, internal constraints, and methodological issues. *Psychological Research*, *86*(8), 2370-2388.

[20]. Bylund, P. L., & Packard, M. D. (2022). Subjective value in entrepreneurship. *Small Business Economics*, *58*(3), 1243-1260.

[21]. Channuntapipat, C., Samsonova-Taddei, A., & Turley, S. (2020). Variation in sustainability assurance practice: An analysis of accounting versus non-accounting providers. *The British Accounting Review*, *52*(2), 100843.

[22]. Chen, M., Tao, Z., Tang, W., Qin, T., Yang, R., & Zhu, C. (2024). Enhancing emergency decision-making with knowledge graphs and large language models. *International Journal of Disaster Risk Reduction*, *113*, 104804.

[23]. Chua, M., Kim, D., Choi, J., Lee, N. G., Deshpande, V., Schwab, J., Lev, M. H., Gonzalez, R. G., Gee, M. S., & Do, S. (2023). Tackling prediction uncertainty in machine learning for healthcare. *Nature Biomedical Engineering*, *7*(6), 711-718.

[24]. De Zarzà, I., De Curtò, J., Roig, G., & Calafate, C. T. (2023). Optimized financial planning: integrating individual and cooperative budgeting models with LLM recommendations. *AI*, *5*(1), 91-114.

[25]. Debnath, R., Darby, S., Bardhan, R., Mohaddes, K., & Sunikka-Blank, M. (2020). Grounded reality meets machine learning: A deep-narrative analysis framework for energy policy research. *Energy research & social science*, *69*, 101704.

[26]. Dove, G., Barca, L., Tummolini, L., & Borghi, A. M. (2022). Words have a weight: Language as a source of inner grounding and flexibility in abstract concepts. *Psychological Research*, *86*(8), 2451-2467.

[27]. Efat Ara, H. (2025). Quantitative Analysis Of Mechanical Testing And Valve Performance In The Oil And Gas Sector: Ensuring Compliance With ISO/IEC 17025 In Global Industrial Infrastructure. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 1424–1457. https://doi.org/10.63125/a5c2w129

[28]. Efat Ara, H., & Shaikh, S. (2023). Hydrogen Embrittlement Sensitivity of Additively Manufactured 347H Stainless Steel: Effects Of Porosity And Residual Stress. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 100–144. https://doi.org/10.63125/kyyasa55

[29]. Erdemir, A., Mulugeta, L., Ku, J. P., Drach, A., Horner, M., Morrison, T. M., Peng, G. C., Vadigepalli, R., Lytton, W. W., & Myers Jr, J. G. (2020). Credible practice of modeling and simulation in healthcare: ten rules from a multidisciplinary perspective. *Journal of translational medicine*, *18*(1), 369.

[30]. Felderer, M., & Ramler, R. (2021). Quality assurance for AI-based systems: Overview and challenges (introduction to interactive session). International Conference on Software Quality,

[31]. Ferdous, W., Manalo, A., Wong, H. S., Abousnina, R., AlAjarmeh, O. S., Zhuge, Y., & Schubel, P. (2020). Optimal design for epoxy polymer concrete based on mechanical properties and durability aspects. *Construction and Building Materials*, *232*, 117229.

[32]. Finlay, C., & Oberman, A. M. (2021). Scaleable input gradient regularization for adversarial robustness. *Machine Learning with Applications*, *3*, 100017.

[33]. Freiesleben, T., & Grote, T. (2023). Beyond generalization: a theory of robustness in machine learning. *Synthese*, *202*(4), 109.

[34]. Freitas, L., Scott III, W. E., & Degenaar, P. (2020). Medicine-by-wire: Practical considerations on formal techniques for dependable medical systems. *Science of Computer Programming*, *200*, 102545.

[35]. Frosolini, A., Catarzi, L., Benedetti, S., Latini, L., Chisci, G., Franz, L., Gennaro, P., & Gabriele, G. (2024). The role of large language models (LLMs) in providing triage for maxillofacial trauma cases: a preliminary study. *Diagnostics*, *14*(8), 839.

[36]. Gandhi, A., & Jain, S. (2020). Adversarial perturbations fool deepfake detectors. 2020 International joint conference on neural networks (IJCNN),

[37]. Ghaffari Laleh, N., Truhn, D., Veldhuizen, G. P., Han, T., van Treeck, M., Buelow, R. D., Langer, R., Dislich, B., Boor, P., & Schulz, V. (2022). Adversarial attacks and adversarial robustness in computational pathology. *Nature communications*, *13*(1), 5711.

[38]. Gholami, H. (2024). Artificial Intelligence Techniques for Sustainable Reconfigurable Manufacturing Systems: An AI-Powered Decision-Making Application Using Large Language Models. *Big Data and Cognitive Computing*, *8*(11), 152.

[39]. Goeuriot, L., Mulhem, P., Quénot, G., Schwab, D., Soulier, L., Di Nunzio, G., Galuščáková, P., de Herrera, A. G. S., Faggioli, G., & Ferro, N. (2024). Experimental IR meets Multilinguality, multimodality, and interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)–Part,

[40]. Gomez-Cabello, C. A., Borna, S., Pressman, S. M., Haider, S. A., & Forte, A. J. (2024). Large language models for intraoperative decision support in plastic surgery: a comparison between ChatGPT-4 and Gemini. *Medicina*, *60*(6), 957.

[41]. Griewing, S., Lechner, F., Gremke, N., Lukac, S., Janni, W., Wallwiener, M., Wagner, U., Hirsch, M., & Kuhn, S. (2024). Proof-of-concept study of a small language model chatbot for breast cancer decision support–a transparent, source-controlled, explainable and data-secure approach. *Journal of Cancer Research and Clinical Oncology*, *150*(10), 451.

[42]. Gumilar, K. E., Indraprasta, B. R., Faridzi, A. S., Wibowo, B. M., Herlambang, A., Rahestyningtyas, E., Irawan, B., Tambunan, Z., Bustomi, A. F., & Brahmantara, B. N. (2024). Assessment of Large Language Models (LLMs) in decision-making support for gynecologic oncology. *Computational and Structural Biotechnology Journal*, *23*, 4019-4026.

[43]. Guo, L. L., Morse, K. E., Aftandilian, C., Steinberg, E., Fries, J., Posada, J., Fleming, S. L., Lemmon, J., Jessa, K., & Shah, N. (2024). Characterizing the limitations of using diagnosis codes in the context of machine learning for healthcare. *BMC Medical Informatics and Decision Making*, *24*(1), 51.

[44]. Habibullah, S. M., & Md. Tahmid Farabe, S. (2022). IOT-Integrated Deep Neural Predictive Maintenance System with Vibration-Signal Diagnostics In Smart Factories. *Journal of Sustainable Development and Policy*, *1*(02), 35-83. https://doi.org/10.63125/6jjq1p95

[45]. Habibullah, S. M., & Muhammad Mohiul, I. (2023). Digital Twin-Driven Thermodynamic and Fluid Dynamic Simulation For Exergy Efficiency In Industrial Power Systems. *American Journal of Scholarly Research and Innovation*, *2*(01), 224–253. https://doi.org/10.63125/k135kt69

[46]. Hagen, A., Jarman, K., Ward, J., Eiden, G., Barinaga, C., Mace, E., Aalseth, C., & Carado, A. (2022). Reduction of detection limit and quantification uncertainty due to interferent by neural classification with abstention. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *1040*, 167174.

[47]. Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., & Kaissis, G. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, *30*(9), 2613-2622.

[48]. Han, L., Ye, H.-J., & Zhan, D.-C. (2024). The capacity and robustness trade-off: Revisiting the channel independent strategy for multivariate time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, *36*(11), 7129-7142.

[49]. Handler, A., Larsen, K. R., & Hackathorn, R. (2024). Large language models present new questions for decision support. *International Journal of Information Management*, *79*, 102811.

[50]. Hansford, H. J., Cashin, A. G., Bagg, M. K., Wewege, M. A., Ferraro, M. C., Kianersi, S., Mayo-Wilson, E., Grant, S. P., Toomey, E., & Skinner, I. W. (2022). Feasibility of an audit and feedback intervention to facilitate journal policy change towards greater promotion of transparency and openness in sports science research. *Sports Medicine-Open*, *8*(1), 101.

[51]. Harvey, L. (2024). Extended Editorial: Defining quality thirty years on: quality, standards, assurance, culture and epistemology. In (Vol. 30, pp. 145-184): Taylor & Francis.

[52]. Heinze-Deml, C., & Meinshausen, N. (2021). Conditional variance penalties and domain shift robustness. *Machine Learning*, *110*(2), 303-348.

[53]. Hendrickx, K., Perini, L., Van der Plas, D., Meert, W., & Davis, J. (2024). Machine learning with a reject option: A survey. *Machine Learning*, *113*(5), 3073-3110.

[54]. Hildesheim, C., & Sonntag, K. (2020). The Quality Culture Inventory: a comprehensive approach towards measuring quality culture in higher education. *Studies in Higher Education*, *45*(4), 892-908.

[55]. Ho, C. N., Tian, T., Ayers, A. T., Aaron, R. E., Phillips, V., Wolf, R. M., Mathioudakis, N., Dai, T., & Klonoff, D. C. (2024). Qualitative metrics from the biomedical literature for evaluating large language models in clinical decision-making: a narrative review. *BMC Medical Informatics and Decision Making*, *24*(1), 357.

[56]. Hobbs, K. L., Mote, M. L., Abate, M. C., Coogan, S. D., & Feron, E. M. (2023). Runtime assurance for safety-critical systems: An introduction to safety filtering approaches for complex control systems. *IEEE Control Systems Magazine*, *43*(2), 28-65.

[57]. Huang, X., Ruan, W., Huang, W., Jin, G., Dong, Y., Wu, C., Bensalem, S., Mu, R., Qi, Y., & Zhao, X. (2024). A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, *57*(7), 175.

[58]. Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457-506.

[59]. Humphrey, C., Sonnerfeldt, A., Komori, N., & Curtis, E. (2021). Audit and the pursuit of dynamic repair. *European Accounting Review*, *30*(3), 445-471.

[60]. Jabed Hasan, T., & Waladur, R. (2023). AI-Driven Cybersecurity, IOT Networking, And Resilience Strategies For Industrial Control Systems: A Systematic Review For U.S. Critical Infrastructure Protection. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 144–176. https://doi.org/10.63125/mbyhj941

[61]. Javed, H., El-Sappagh, S., & Abuhmed, T. (2024). Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. *Artificial Intelligence Review*, *58*(1), 12.

[62]. Jia, Y., McDermid, J., Lawton, T., & Habli, I. (2022). The role of explainability in assuring safety of machine learning in healthcare. *IEEE Transactions on Emerging Topics in Computing*, *10*(4), 1746-1760.

[63]. Jinnat, A. (2025). Machine-Learning Models For Predicting Blood Pressure And Cardiac Function Using Wearable Sensor Data. *International Journal of Scientific Interdisciplinary Research*, *6*(2), 102–142. https://doi.org/10.63125/h7rbyt25

[64]. Jinnat, A., & Md. Kamrul, K. (2021). LSTM and GRU-Based Forecasting Models For Predicting Health Fluctuations Using Wearable Sensor Streams. *American Journal of Interdisciplinary Studies*, *2*(02), 32-66. https://doi.org/10.63125/1p8gbp15

[65]. Kompa, B., Snoek, J., & Beam, A. L. (2021). Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine*, *4*(1), 4.

[66]. Kotla, B., & Bosman, L. (2023). Redefining sustainability and entrepreneurship teaching. *Trends in Higher Education*, *2*(3), 498-513.

[67]. Kuang, X., Ma, C., & Ren, Y.-S. (2024). Credit risk: A new privacy-preserving decentralized credit assessment model. *Finance Research Letters*, *67*, 105937.

[68]. Kumar, M., & Chand, S. (2020). A secure and efficient cloud-centric internet-of-medical-things-enabled smart healthcare system with public verifiability. *IEEE Internet of Things Journal*, *7*(10), 10650-10659.

[69]. Kumar, M., & Chand, S. (2021). A provable secure and lightweight smart healthcare cyber-physical system with public verifiability. *IEEE Systems Journal*, *16*(4), 5501-5508.

[70]. Kuznetsov, O., Rusnak, A., Yezhov, A., Kanonik, D., Kuznetsova, K., & Karashchuk, S. (2024). Enhanced security and efficiency in blockchain with aggregated zero-knowledge proof mechanisms. *IEEE access*, *12*, 49228-49248.

[71]. Labbadi, M., & Cherkaoui, M. (2021). Adaptive fractional-order nonsingular fast terminal sliding mode based robust tracking control of quadrotor UAV with Gaussian random disturbances and uncertainties. *IEEE Transactions on Aerospace and Electronic Systems*, *57*(4), 2265-2277.

[72]. Lambert, B., Forbes, F., Doyle, S., Dehaene, H., & Dojat, M. (2024). Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis. *Artif. Intell. Medicine*, *150*, 102830.

[73]. Lan, Y., Wu, Y., Xu, W., Feng, W., & Zhang, Y. (2024). Chinese fine-grained financial sentiment analysis with large language models. *Neural Computing and Applications*, 1-10.

[74]. Lawson McLean, A., Wu, Y., Lawson McLean, A. C., & Hristidis, V. (2024). Large language models as decision aids in neuro-oncology: a review of shared decision-making applications. *Journal of Cancer Research and Clinical Oncology*, *150*(3), 139.

[75]. Le, W. T., Vorontsov, E., Romero, F. P., Seddik, L., Elsharief, M. M., Nguyen-Tan, P. F., Roberge, D., Bahig, H., & Kadoury, S. (2022). Cross-institutional outcome prediction for head and neck cancer patients using self-attention neural networks. *Scientific reports*, *12*(1), 3183.

[76]. Lear, M. K., Spata, A., Tittler, M., Fishbein, J. N., Arch, J. J., & Luoma, J. B. (2023). Transparency and reproducibility in the journal of contextual behavioral science: An audit study. *Journal of Contextual Behavioral Science*, *28*, 207-214.

[77]. Li, Z., Nagrebetsky, A., Ranjeva, S., Bi, N., Liu, D., Vidal Melo, M. F., Houle, T., Yin, L., & Deng, H. (2023). A transformer-based deep learning algorithm to auto-record undocumented clinical one-lung ventilation events. International Workshop on Health Intelligence,

[78]. Liu, A., Liu, X., Yu, H., Zhang, C., Liu, Q., & Tao, D. (2021). Training robust deep neural networks via adversarial noise propagation. *IEEE Transactions on Image Processing*, *30*, 5769-5781.

[79]. Liu, Z., Kou, J., Zhang, W., Gu, C., Fang, X., Huang, Z., Yuan, H., Li, H., Lu, X., & Yin, A. (2024). Comprehensive Evaluation of AI Hallucination and Novel UV-Oriented Framework toward Safe and Trustworthy AI. 2024 7th International Conference on Universal Village (UV),

[80]. Luo, M., Gokhale, T., Varshney, N., Yang, Y., & Baral, C. (2024). *Advances in Multimodal Information Retrieval and Generation*. Springer.

[81]. Mac Donald, K., Rezania, D., & Baker, R. (2020). A grounded theory examination of project managers' accountability. *International Journal of Project Management*, *38*(1), 27-35.

[82]. Macleod, M., & Group, U. o. E. R. S. (2022). Improving the reproducibility and integrity of research: what can different stakeholders contribute? *BMC Research Notes*, *15*(1), 146.

[83]. Mahmoud, A., Aggarwal, N., Nobbe, A., Vicarte, J. R. S., Adve, S. V., Fletcher, C. W., Frosio, I., & Hari, S. K. S. (2020). Pytorchfi: A runtime perturbation tool for dnns. 2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W),

[84]. Mao, C., Gupta, A., Nitin, V., Ray, B., Song, S., Yang, J., & Vondrick, C. (2020). Multitask learning strengthens adversarial robustness. European Conference on Computer Vision,

[85]. Md Arman, H., & Md Nahid, H. (2023). The Influence Of IOT And Digital Technologies On Financial Risk Monitoring And Investment Efficiency In Global Supply Chains. *American Journal of Interdisciplinary Studies*, 4(02), 91-125. https://doi.org/10.63125/e6yt5x19

[86]. Md Arman, H., & Md.Kamrul, K. (2022). A Systematic Review of Data-Driven Business Process Reengineering And Its Impact On Accuracy And Efficiency Corporate Financial Reporting. *International Journal of Business and Economics Insights*, *2*(4), 01–41. https://doi.org/10.63125/btx52a36

[87]. Md Harun-Or-Rashid, M. (2024). Blockchain Adoption And Organizational Long-Term Growth In Small And Medium Enterprises (SMEs). *Review of Applied Science and Technology*, *3*(04), 128–164. https://doi.org/10.63125/rq0zds79

[88]. Md Harun-Or-Rashid, M. (2025a). AI-Driven Threat Detection and Response Framework For Cloud Infrastructure Security. *American Journal of Scholarly Research and Innovation*, *4*(01), 494–535. https://doi.org/10.63125/e58hzh78

[89]. Md Harun-Or-Rashid, M. (2025b). Is The Metaverse the Next Frontier for Corporate Growth And Innovation? Exploring The Potential of The Enterprise Metaverse. *American Journal of Interdisciplinary Studies*, *6*(1), 354-393. https://doi.org/10.63125/ckd54306

[90]. Md Harun-Or-Rashid, M., & Sai Praveen, K. (2022). Data-Driven Approaches To Enhancing Human–Machine Collaboration In Remote Work Environments. *International Journal of Business and Economics Insights*, *2*(3), 47-83. https://doi.org/10.63125/wt9t6w68

[91]. Md, K., & Sai Praveen, K. (2024). Hybrid Discrete-Event And Agent-Based Simulation Framework (H-DEABSF) For Dynamic Process Control In Smart Factories. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *4*(1), 72–96. https://doi.org/10.63125/wcqq7x08

[92]. Md Mesbaul, H. (2023). A Meta-Analysis of Lean Merchandising Strategies In Fashion Retail: Global Insights From The Post-Pandemic Era. *Review of Applied Science and Technology*, *2*(04), 94-123. https://doi.org/10.63125/y8x4k683

[93]. Md Mesbaul, H. (2025). A Framework-Based Meta-Analysis Of Artificial Intelligence-Driven ERP Solutions For Circular And Sustainable Supply Chains. *International Journal of Scientific Interdisciplinary Research*, *6*(1), 327-367. https://doi.org/10.63125/n6k7r711

[94]. Md Milon, M., & Md. Mominul, H. (2023). The Impact Of Bim And Digital Twin Technologies On Risk Reduction In Civil Infrastructure Projects. *American Journal of Advanced Technology and Engineering Solutions*, *3*(04), 01-41. https://doi.org/10.63125/xgyzqk40

[95]. Md Mohaiminul, H., & Alifa Majumder, N. (2024). Deep Learning And Graph Neural Networks For Real-Time Cybersecurity Threat Detection. *Review of Applied Science and Technology*, *3*(01), 106–142. https://doi.org/10.63125/dp38xp64

[96]. Md Mohaiminul, H., & Md Muzahidul, I. (2023). Reinforcement Learning Approaches to Optimize IT Service Management Under Data Security Constraints. *American Journal of Scholarly Research and Innovation*, *2*(02), 373-414. https://doi.org/10.63125/z7q4cy92

[97]. Md Musfiqur, R., & Md.Kamrul, K. (2023). Mechanisms By Which AI-Enabled CRM Systems Influence Customer Retention and Overall Business Performance: A Systematic Literature Review Of Empirical Findings. *International Journal of Business and Economics Insights*, *3*(1), 31-67. https://doi.org/10.63125/qqe2bm11

[98]. Md Rezaul, K., & Md.Kamrul, K. (2023). Integrating AI-Powered Robotics in Large-Scale Warehouse Management: Enhancing Operational Efficiency, Cost Reduction, And Supply Chain Performance Models. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 01-30. https://doi.org/10.63125/mszb5c17

[99]. Md. Al Amin, K., & Sai Praveen, K. (2023). The Role of Industrial Engineering In Advancing Sustainable Manufacturing And Quality Compliance In Global Engineering Systems. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 31–61. https://doi.org/10.63125/8w1vk676

[100]. Md. Foysal, H., & Abdulla, M. (2024). Agile And Sustainable Supply Chain Management Through AI-Based Predictive Analytics And Digital Twin Simulation. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 343–376. https://doi.org/10.63125/sejyk977

[101]. Md. Hasan, I., & Shaikat, B. (2021). Global Sourcing, Cybersecurity Vulnerabilities, And U.S. Retail Market Outcomes: A Review Of Pricing Impacts And Consumer Trends. *American Journal of Scholarly Research and Innovation*, *1*(01), 126–166. https://doi.org/10.63125/78jcs795

[102]. Md. Jobayer Ibne, S., & Aditya, D. (2024). Machine Learning and Secure Data Pipeline Frameworks For Improving Patient Safety Within U.S. Electronic Health Record Systems. *American Journal of Interdisciplinary Studies*, *5*(03), 43–85. https://doi.org/10.63125/nb2c1f86

[103]. Md. Milon, M. (2025). A Review On The Influence Of AI-Enabled Fire Detection And Suppression Systems In Enhancing Building Safety. *Review of Applied Science and Technology*, *4*(04), 36–73. https://doi.org/10.63125/h0dbee62

[104]. Md. Milon, M., & Md. Mominul, H. (2024). Quantitative Assessment Of Hydraulic Modeling Tools In Optimizing Fire Sprinkler System Efficiency. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 415–448. https://doi.org/10.63125/6dsw5w30

[105]. Md. Mosheur, R. (2025). AI-Driven Predictive Analytics Models For Enhancing Group Insurance Portfolio Performance And Risk Forecasting. *International Journal of Scientific Interdisciplinary Research*, *6*(2), 39–87. https://doi.org/10.63125/qh5qgk22

[106]. Md. Mosheur, R., & Md Arman, H. (2024). Impact Of Big Data and Predictive Analytics On Financial Forecasting Accuracy And Decision-Making In Global Capital Markets. *American Journal of Scholarly Research and Innovation*, *3*(02), 99–140. https://doi.org/10.63125/hg37h121

[107]. Md. Rabiul, K. (2025). Artificial Intelligence-Enhanced Predictive Analytics For Demand Forecasting In U.S. Retail Supply Chains. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 959–993. https://doi.org/10.63125/gbkf5c16

[108]. Md. Rabiul, K., & Mohammad Mushfequr, R. (2023). A Quantitative Study On Erp-Integrated Decision Support Systems In Healthcare Logistics. *Review of Applied Science and Technology*, *2*(01), 142–184. https://doi.org/10.63125/c92bbj37

[109]. Md. Rabiul, K., & Samia, A. (2021). Integration Of Machine Learning Models And Advanced Computing For Reducing Logistics Delays In Pharmaceutical Distribution. *American Journal of Advanced Technology and Engineering Solutions*, *1*(4), 01-42. https://doi.org/10.63125/ahnkqj11

[110]. Md.Kamrul, K., & Md Omar, F. (2022). Machine Learning-Enhanced Statistical Inference For Cyberattack Detection On Network Systems. *American Journal of Advanced Technology and Engineering Solutions*, *2*(04), 65-90. https://doi.org/10.63125/sw7jzx60

[111]. Mechali, O., Xu, L., Huang, Y., Shi, M., & Xie, X. (2021). Observer-based fixed-time continuous nonsingular terminal sliding mode control of quadrotor aircraft under uncertainties and disturbances for robust trajectory tracking: Theory and experiment. *Control Engineering Practice*, *111*, 104806.

[112]. Mehandru, N., Miao, B. Y., Almaraz, E. R., Sushil, M., Butte, A. J., & Alaa, A. (2024). Evaluating large language models as agents in the clinic. *NPJ Digital Medicine*, *7*(1), 84.

[113]. Meng, M. H., Bai, G., Teo, S. G., Hou, Z., Xiao, Y., Lin, Y., & Dong, J. S. (2022). Adversarial robustness of deep neural networks: A survey from a formal verification perspective. *IEEE Transactions on Dependable and Secure Computing*.

[114]. Mirzaei, T., Amini, L., & Esmaeilzadeh, P. (2024). Clinician voices on ethics of LLM integration in healthcare: a thematic analysis of ethical concerns and implications. *BMC Medical Informatics and Decision Making*, *24*(1), 250.

[115]. Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a service: a pragmatic operationalisation of AI ethics. *Minds and Machines*, *31*(2), 239-256.

[116]. Mst. Shahrin, S. (2025). Predictive Neural Network Models For Cyberattack Pattern Recognition And Critical Infrastructure Vulnerability Assessment. *Review of Applied Science and Technology*, *4*(02), 777-819. https://doi.org/10.63125/qp0de852

[117]. Mst. Shahrin, S., & Samia, A. (2023). High-Performance Computing For Scaling Large-Scale Language And Data Models In Enterprise Applications. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *3*(1), 94–131. https://doi.org/10.63125/e7yfwm87

[118]. Mubarak, R., Alsboui, T., Alshaikh, O., Inuwa-Dutse, I., Khan, S., & Parkinson, S. (2023). A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. *IEEE access*, *11*, 144497-144529.

[119]. Muhammad Mohiul, I. (2020). Impact Of Digital Construction Management Platforms on Project Performance Post-Covid-19. *American Journal of Interdisciplinary Studies*, *1*(04), 01-25. https://doi.org/10.63125/nqp0zh08

[120]. Muhammad Mohiul, I., & Rahman, M. D. H. (2021). Quantum-Enhanced Charge Transport Modeling In Perovskite Solar Cells Using Non-Equilibrium Green's Function (NEGF) Framework. *Review of Applied Science and Technology*, *6*(1), 230–262. https://doi.org/10.63125/tdbjaj79

[121]. Natarajan, R., Lokesh, G. H., Flammini, F., Premkumar, A., Venkatesan, V. K., & Gupta, S. K. (2023). A novel framework on security and energy enhancement based on internet of medical things for healthcare 5.0. *Infrastructures*, *8*(2), 22.

[122]. Nielsen, I. E., Dera, D., Rasool, G., Ramachandran, R. P., & Bouaynaya, N. C. (2022). Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine*, *39*(4), 73-84.

[123]. Oniani, D., Wu, X., Visweswaran, S., Kapoor, S., Kooragayalu, S., Polanska, K., & Wang, Y. (2024). Enhancing large language models for clinical decision support by incorporating clinical practice guidelines. 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI),

[124]. Ostblom, J., & Timbers, T. (2022). Opinionated practices for teaching reproducibility: motivation, guided instruction and practice. *Journal of Statistics and Data Science Education*, *30*(3), 241-250.

[125]. Pandey, P., Chasmai, M., Sur, T., & Lall, B. (2023). Robust prototypical few-shot organ segmentation with regularized neural-odes. *IEEE Transactions on Medical Imaging*, *42*(9), 2490-2501.

[126]. Pankaz Roy, S. (2023). Epidemiological Trends In Zoonotic Diseases Comparative Insights From South Asia And The U.S. *American Journal of Interdisciplinary Studies*, *4*(03), 166–207. https://doi.org/10.63125/wrrfmt97

[127]. Park, Y.-J., Pillai, A., Deng, J., Guo, E., Gupta, M., Paget, M., & Naugler, C. (2024). Assessing the research landscape and clinical utility of large language models: a scoping review. *BMC Medical Informatics and Decision Making*, *24*(1), 72.

[128]. Patel, K., Beluch, W., Rambach, K., Cozma, A.-E., Pfeiffer, M., & Yang, B. (2021). Investigation of uncertainty of deep learning-based object classification on radar spectra. 2021 IEEE Radar Conference (RadarConf21),

[129]. Pecorelli, F., Catolino, G., Ferrucci, F., De Lucia, A., & Palomba, F. (2022). Software testing and android applications: a large-scale empirical study. *Empirical Software Engineering*, *27*(2), 31.

[130]. Petrillo, L., Martinelli, F., Santone, A., & Mercaldo, F. (2024). Toward the adoption of explainable pre-trained large language models for classifying human-written and ai-generated sentences. *Electronics*, *13*(20), 4057.

[131]. Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2020). Secure and robust machine learning for healthcare: A survey. *IEEE Reviews in Biomedical Engineering*, *14*, 156-180.

[132]. Rahman, M. D. H. (2022). Modelling The Impact Of Temperature Coefficients On PV System Performance In Hot And Humid Climates. *International Journal of Scientific Interdisciplinary Research*, *1*(01), 194–237. https://doi.org/10.63125/abj6wy92

[133]. Rahman, S. M. T., & Abdul, H. (2021). The Role Of Predictive Analytics In Enhancing Agribusiness Supply Chains. *Review of Applied Science and Technology*, *6*(1), 183–229. https://doi.org/10.63125/n9z10h68

[134]. Rahman, S. M. T., & Aditya, D. (2024). Market-Driven Management Strategies Using Artificial Intelligence To Strengthen Food Safety And Advance One Health Initiatives. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 377–414. https://doi.org/10.63125/0f9wah05

[135]. Raiaan, M. A. K., Mukta, M. S. H., Fatema, K., Fahad, N. M., Sakib, S., Mim, M. M. J., Ahmad, J., Ali, M. E., & Azam, S. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, *12*, 26839-26874.

[136]. Rakibul, H. (2025). A Systematic Review Of Human-AI Collaboration In It Support Services: Enhancing User Experience And Workflow Automation. *American Journal of Interdisciplinary Studies*, *6*(3), 01-37. https://doi.org/10.63125/0fd1yb74

[137]. Rakibul, H., & Alifa Majumder, N. (2023). AI Applications In Emerging Tech Sectors: A Review Of AI Use Cases Across Healthcare, Retail, And Cybersecurity. *American Journal of Scholarly Research and Innovation*, *2*(02), 336–372. https://doi.org/10.63125/adtgfj55

[138]. Rauniyar, A., Hagos, D. H., Jha, D., Håkegård, J. E., Bagci, U., Rawat, D. B., & Vlassov, V. (2023). Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions. *IEEE Internet of Things Journal*, *11*(5), 7374-7398.

[139]. Ren, C., & Xu, Y. (2022). Robustness verification for machine-learning-based power system dynamic security assessment models under adversarial examples. *IEEE Transactions on Control of Network Systems*, *9*(4), 1645-1654.

[140]. Revell, T., Yeadon, W., Cahilly-Bretzin, G., Clarke, I., Manning, G., Jones, J., Mulley, C., Pascual, R., Bradley, N., & Thomas, D. (2024). ChatGPT versus human essayists: an exploration of the impact of artificial intelligence for authorship and academic integrity in the humanities. *International Journal for Educational Integrity*, *20*(1), 18.

[141]. Rifat, C., & Rebeka, S. (2023). The Role Of ERP-Integrated Decision Support Systems In Enhancing Efficiency And Coordination In Healthcare Logistics: A Quantitative Study. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 265–285. https://doi.org/10.63125/c7srk144

[142]. Rony, M. A., & Samia, A. (2022). Digital Twin Frameworks for Enhancing Climate-Resilient Infrastructure Design. *Review of Applied Science and Technology*, *1*(01), 38–70. https://doi.org/10.63125/54zej644

[143]. Rusak, E., Schott, L., Zimmermann, R. S., Bitterwolf, J., Bringmann, O., Bethge, M., & Brendel, W. (2020). A simple way to make neural networks robust against diverse image corruptions. European conference on computer vision,

[144]. Saba, A., & Md. Sakib Hasan, H. (2024). Machine Learning And Secure Data Pipelines For Enhancing Patient Safety In Electronic Health Record (EHR) Among U.S. Healthcare Providers. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *4*(1), 124–168. https://doi.org/10.63125/qm4he747

[145]. Sabuj Kumar, S. (2023). Integrating Industrial Engineering and Petroleum Systems With Linear Programming Model For Fuel Efficiency And Downtime Reduction. *Journal of Sustainable Development and Policy*, *2*(04), 108-139. https://doi.org/10.63125/v7d6a941

[146]. Sabuj Kumar, S. (2024). Petroleum Storage Tank Design and Inspection Using Finite Element Analysis Model For Ensuring Safety Reliability And Sustainability. *Review of Applied Science and Technology*, *3*(04), 94–127. https://doi.org/10.63125/a18zw719

[147]. Sabuj Kumar, S. (2025). AI Driven Predictive Maintenance In Petroleum And Power Systems Using Random Forest Regression Model For Reliability Engineering Framework. *American Journal of Scholarly Research and Innovation*, *4*(01), 363-391. https://doi.org/10.63125/477x5t65

[148]. Sai Praveen, K. (2024). AI-Enhanced Data Science Approaches For Optimizing User Engagement In U.S. Digital Marketing Campaigns. *Journal of Sustainable Development and Policy*, *3*(03), 01-43. https://doi.org/10.63125/65ebsn47

[149]. Sai Praveen, K., & Md, K. (2025). Real-Time Cyber-Physical Deployment and Validation Of H-DEABSF: Model Predictive Control, And Digital-Twin–Driven Process Control In Smart Factories. *Review of Applied Science and Technology*, *4*(02), 750-776. https://doi.org/10.63125/yrkm0057

[150]. Saied, W. M., Elakhdar, B. E., & Hassan, D. G. (2024). Comprehensive synthesis of decision-making in complex systems. 2024 6th International Conference on Computing and Informatics (ICCI),

[151]. Saikat, S., & Aditya, D. (2023). Reliability-Centered Maintenance Optimization Using Multi-Objective Ai Algorithms In Refinery Equipment. *American Journal of Scholarly Research and Innovation*, *2*(01), 389–411. https://doi.org/10.63125/6a6kqm73

[152]. Salim, S., & Jayasudha, J. (2023). A literature survey on estimating uncertainty in deep learning models: Ensuring safety in intelligent systems. 2023 2nd International Conference on Computational Systems and Communication (ICCSC),

[153]. Sandmann, S., Riepenhausen, S., Plagwitz, L., & Varghese, J. (2024). Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. *Nature communications*, *15*(1), 2050.

[154]. Sangeetha, S., Selvarathi, C., Mathivanan, S. K., Cho, J., & Easwaramoorthy, S. V. (2024). Secure healthcare access control system (SHACS) for anomaly detection and enhanced security in cloud-based healthcare applications. *IEEE access*, *12*, 164543-164559.

[155]. Shaikat, B., & Aditya, D. (2024). Graph Neural Network Models For Predicting Cyber Attack Patterns In Critical Infrastructure Systems. *Review of Applied Science and Technology*, *3*(01), 68–105. https://doi.org/10.63125/pmnqxk63

[156]. Sharmin, S., Rathi, N., Panda, P., & Roy, K. (2020). Inherent adversarial robustness of deep spiking neural networks: Effects of discrete input encoding and non-linear activations. European Conference on Computer Vision,

[157]. Sim, J. J., Loh, S. H., Wong, K. L., & Choong, C. K. (2021). Do we need trust transfer mechanisms? An M-commerce adoption perspective. *Journal of Theoretical and Applied Electronic Commerce Research*, *16*(6), 2241-2262.

[158]. Steenhoek, B., Rahman, M. M., Jiles, R., & Le, W. (2023). An empirical study of deep learning models for vulnerability detection. 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE),

[159]. Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., & Müller, K.-R. (2021). Towards CRISP-ML (Q): a machine learning process model with quality assurance methodology. *Machine learning and knowledge extraction*, *3*(2), 392-413.

[160]. Sumi, R. S., & Kabir, G. (2021). Satisfaction of e-learners with electronic learning service quality using the servqual model. *Journal of Open Innovation: Technology, Market, and Complexity*, *7*(4), 227.

[161]. Sun, X., & Sun, S. (2021). Adversarial robustness and attacks for multi-view deep models. *Engineering Applications of Artificial Intelligence*, *97*, 104085.

[162]. Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., & Fu, S. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *NPJ Digital Medicine*, *7*(1), 258.

[163]. Tambon, F., Laberge, G., An, L., Nikanjam, A., Mindom, P. S. N., Pequignot, Y., Khomh, F., Antoniol, G., Merlo, E., & Laviolette, F. (2022). How to certify machine learning based safety-critical systems? A systematic literature review. *Automated Software Engineering*, *29*(2), 38.

[164]. Taylor, D., Park, Y. S., Smith, C., Cate, O. t., & Tekian, A. (2020). Constructing approaches to entrustable professional activity development that deliver valid descriptions of professional practice. *Teaching and Learning in Medicine*, *33*(1), 89-97.

[165]. Thulasidasan, S., Thapa, S., Dhaubhadel, S., Chennupati, G., Bhattacharya, T., & Bilmes, J. (2021). An effective baseline for robustness to distributional shift. 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA),

[166]. Tian, Y., Liu, Y., Pang, G., Liu, F., Chen, Y., & Carneiro, G. (2022). Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. European Conference on Computer Vision,

[167]. Tupayachi, J., Xu, H., Omitaomu, O. A., Camur, M. C., Sharmin, A., & Li, X. (2024). Towards next-generation urban decision support systems through ai-powered construction of scientific ontology using large language models—a case in optimizing intermodal freight transportation. *Smart Cities*, *7*(5), 2392-2421.

[168]. Tyurin, A. A., Suhorukova, A. V., Kabardaeva, K. V., & Goldenkova-Pavlova, I. V. (2020). Transient gene expression is an effective experimental tool for the research into the fine mechanisms of plant gene function: advantages, limitations, and solutions. *Plants*, *9*(9), 1187.

[169]. Ullah, E., Parwani, A., Baig, M. M., & Singh, R. (2024). Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology*, *19*(1), 43.

[170]. Vueghs, C., Shakeri, H., Renton, T., & Van der Cruyssen, F. (2024). Development and evaluation of a GPT4-based orofacial pain clinical decision support system. *Diagnostics*, *14*(24), 2835.

[171]. Wang, M., Chen, Y., & Yan, J. (2023). Credit Default Prediction Model based on Horizontal Federated Neural Network and Improved TrAdaBoost Algorithm. 2023 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics),

[172]. Wang, Z., Liang, Y., Sun, W., Xu, C., Zhou, Y., & Zhang, Y. (2024). Initial-LLM: A Large Language Model-Guided Metaheuristic Framework for Enhanced Feature Selection in Clinical Decision Support Systems. International Conference on Health Information Science,

[173]. Weber, M., Xu, X., Karlaš, B., Zhang, C., & Li, B. (2023). Rab: Provable robustness against backdoor attacks. 2023 IEEE Symposium on Security and Privacy (SP),

[174]. Wiesenfeld, B. M., Aphinyanaphongs, Y., & Nov, O. (2022). AI model transferability in healthcare: a sociotechnical perspective. *Nature Machine Intelligence*, *4*(10), 807-809.

[175]. Williams, C. Y., Miao, B. Y., Kornblith, A. E., & Butte, A. J. (2024). Evaluating the use of large language models to provide clinical recommendations in the Emergency Department. *Nature communications*, *15*(1), 8236.

[176]. Zamal Haider, S., & Mst. Shahrin, S. (2021). Impact Of High-Performance Computing In The Development Of Resilient Cyber Defense Architectures. *American Journal of Scholarly Research and Innovation*, *1*(01), 93–125. https://doi.org/10.63125/fradxg14

[177]. Zeng, X., Linwood, S. L., & Liu, C. (2022). Pretrained transformer framework on pediatric claims data for population specific tasks. *Scientific reports*, *12*(1), 3651.

[178]. Zhang, C., Liu, A., Liu, X., Xu, Y., Yu, H., Ma, Y., & Li, T. (2020). Interpreting and improving adversarial robustness of deep neural networks with neuron sensitivity. *IEEE Transactions on Image Processing*, *30*, 1291-1304.

[179]. Zhang, C., Yang, Z., He, X., & Deng, L. (2020). Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, *14*(3), 478-493.

[180]. Zhang, J., Li, H., Xu, D., Lou, Y., Ran, M., Jin, Z., & Huang, Y. (2024). Decouple and Decorrelate: A Disentanglement Security Framework Combining Sample Weighting for Cross-Institution Biased Disease Diagnosis. *IEEE Internet of Things Journal*, *11*(15), 25543-25557.

[181]. Zhang, Y., & Ng, S. T. (2021). A hypothesis-driven framework for resilience analysis of public transport network under compound failure scenarios. *International Journal of Critical Infrastructure Protection*, *35*, 100455.

[182]. Zhong, X., & Liu, R. (2024). Robustness analysis of large scientific facilities development network with different cascading failure modes. *Computers & Industrial Engineering*, *193*, 110281.

[183]. Zhou, W., Zhu, X., Han, Q.-L., Li, L., Chen, X., Wen, S., & Xiang, Y. (2024). The security of using large language models: A survey with emphasis on ChatGPT. *IEEE/CAA Journal of Automatica Sinica*.

[184]. Zhu, L., Rong, Y., McGee, L. A., Rwigema, J.-C. M., & Patel, S. H. (2024). Testing and validation of a custom retrained large language model for the supportive care of HN patients with external knowledge base. *Cancers*, *16*(13), 2311.

[185]. Zulqarnain, F. N. U., & Subrato, S. (2021). Modeling Clean-Energy Governance Through Data-Intensive Computing And Smart Forecasting Systems. *International Journal of Scientific Interdisciplinary Research*, *2*(2), 128–167. https://doi.org/10.63125/wnd6qs51

[186]. Zulqarnain, F. N. U., & Subrato, S. (2023). Intelligent Climate Risk Modeling For Robust Energy Resilience And National Security. *Journal of Sustainable Development and Policy*, *2*(04), 218-256. https://doi.org/10.63125/jmer2r39