UtilitasMathematica
ISSN 0315-3681 Volume 122 (2), 2025

A Zero Trust–Guided Safeguarding Framework for Generative AI Systems in High-Consequence Environments

Dr. Aniket Satish Deshpande

Researcher, Department of Computer Science and Engineering, Sunrise University, Alwar, India

Email ID: anik.deshpande@gmail.com

Abstract:

Generative AI is moving into everyday decision-making work in places like hospitals, trading floors, public service offices, and industrial monitoring rooms. It is no longer limited to drafting text or summarizing documents. In many cases, the output of a model becomes part of how a situation is understood, which can influence what a professional decides to do next. This influence is subtle. A system does not need to be wrong to shift judgment; it only needs to present information with a certain tone or emphasis. That is where risks begin to appear, especially when prompt history, retrieval sources, or context can be nudged in ways that are hard to detect afterward.

Traditional cybersecurity controls focus on who is allowed to access a system and whether the system is functioning as intended. That is necessary, but in high-consequence environments, it is not enough. The question is whether the reasoning the model produces should be allowed to guide or inform action.

This paper introduces the Zero Trust–Guided Generative AI Cybersecurity Safeguarding Maturity Model (ZT-GAI-CSMM), which applies continuous verification at the point where model output intersects with operational decisions. Case applications in finance, clinical care, and industrial control settings demonstrate how organizations can gain value from Generative AI while maintaining accountability and oversight.

Keywords: Zero Trust; Generative AI; Cybersecurity Governance; AI Risk Management; High-Consequence Systems; Interpretive Risk; Model Influence Control; Retrieval-Augmented Generation (RAG) Security; Context Integrity Validation; Cognitive Interaction Governance; Operational Action Boundary Control; Human-in-the-Loop (HITL) Oversight; Decision Accountability; Drift Detection; AI Assurance Frameworks

1. Introduction

Generative AI is now showing up in routine decision-making work in places where outcomes are costly to get wrong. A doctor may glance at a generated clinical summary while reviewing patient history; a credit officer might read a model-produced narrative when considering a borderline loan case; a control room operator may see an AI-generated explanation alongside sensor deviations [1], [2]. The model isn't just answering questions. It is shaping how the situation is *understood*. A slight change in tone or emphasis can make one option feel more reasonable than another, even when the underlying evidence is the same [3]. The effect is subtle, which is exactly why it is easy to overlook.

This influence is not something we encountered with earlier predictive models. Those systems could be tested against benchmarks and expected outcomes [4], [5]. Generative systems produce explanations, not just outputs. Two responses based on the same data may highlight different factors but still sound coherent. In environments where decisions must be explained or defended later, those differences in framing matter more than we usually admit [6].

Most cybersecurity and governance controls focus on identity, access, and model validation. Necessary, yes — but they don't help once the text is in front of a human decision-maker [7], [8]. A model can be properly deployed and still shift judgment if prompt history, retrieval sources, or contextual memory shift over time.

Zero Trust principles help by removing assumptions of default trust and requiring ongoing verification [9], [10]. But Zero Trust, as usually implemented, stops at the moment the output is delivered. It does not examine whether that output should influence an operational decision.

This paper introduces the Zero Trust–Guided Generative AI Safeguarding Maturity Model (ZT-GAI-CSMM). The model focuses on the *moment of influence* — where inference shapes action. The aim is not to limit the use of Generative AI, but to ensure that when it does influence decisions, that influence is visible, accountable, and aligned with organizational risk tolerance.

2.0 Related Work and Limitations of Existing Frameworks

Recent AI governance efforts have focused on transparency, dataset documentation, privacy controls, and structured impact assessments. Examples include the NIST AI Risk Management Framework, the OECD AI Principles, the EU AI Act, and ISO/IEC 42001 for AI management systems [1]–[4]. These frameworks help organizations define oversight responsibilities and deployment safeguards. However, they are built on assumptions carried over from earlier predictive and classification models, where outputs can be measured against reference data or known targets. Generative systems behave differently. They produce open-ended, context-dependent text that does not have a single correct answer, making many traditional assurance measures less effective.

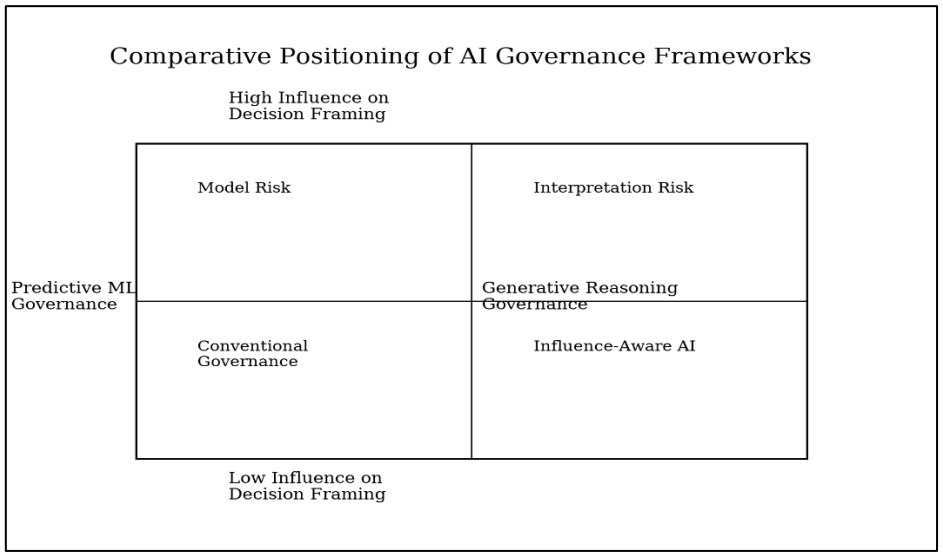

Figure 1. Comparative positioning of governance approaches ranging from performance-centered predictive ML controls to interpretive influence controls required for Generative AI.

Model Risk Management practices in finance and infrastructure typically rely on deviations from expected performance to classify and control risk [5], [6]. With Generative AI, the concern is not simply statistical performance. It is the influence the model has on how a situation is interpreted. A response may appear coherent and convincing while resting on weak or unsupported reasoning [7], [8]. These effects often pass through accuracy checks unnoticed.

Zero Trust architectures address identity verification and access boundaries, and they limit lateral movement and credential misuse [9], [10]. But they do not examine whether the output a user receives is trustworthy, justified, or appropriate for the decision at hand. A fully authenticated user can still be influenced by reasoning that is confidently stated but poorly grounded.

Work in explainability and interpretability seeks to make model behavior more transparent. Yet most techniques provide post-hoc descriptions rather than access to the actual reasoning process [11], [12]. In generative systems, explanations may mimic the same persuasive tone as the original output, increasing confidence without increasing correctness.

Content moderation and hallucination-reduction strategies typically focus on blocking harmful, biased, or factually incorrect statements [13], [14]. High-consequence failures rarely present in this form. They occur when a model produces plausible but subtly misleading reasoning — for example, suggesting urgency where none exists, or implying causal links that are not established [15], [16].

Retrieval-Augmented Generation (RAG) improves factual grounding by linking output to external sources [17], [18]. However, it does not guarantee that the retrieved evidence is interpreted correctly, nor does it prevent context-poisoning attacks in which an attacker introduces manipulated documents into retrieval systems [19], [20].

A growing body of research suggests that failures in Generative AI are primarily **epistemic** rather than computational [21]. The risk lies in how reasoning shapes judgment, not merely whether the output contains errors. Existing governance frameworks address the output, but not the moment when output becomes action. That gap requires safeguards that continuously evaluate provenance, reasoning stability, and the consequence of acting on the model's response, rather than assuming that fluency or system access controls imply trustworthiness.

3.0 Generative AI Threat Surface Expansion

When Generative AI is brought into operational work, the risk does not look like the traditional "find-and-patch" vulnerability picture. The system starts influencing how people *make sense* of what they see. That is the key shift. A model can present information in a way that nudges interpretation, even when all the facts are technically correct [22]. So the failure does not show up as a wrong answer. It shows up as a changed judgment, which is much harder to detect.

3.1 Interpretive and Cognitive Influence

Generative AI has a way of sounding sure of itself. The tone is smooth, organized, and almost clinical. That alone can make the output feel trustworthy. In practice, this matters more than it seems. Consider a nurse deciding whether to escalate a case, or an analyst sorting financial risk alerts, or an engineer watching pressure and temperature readings in a control room. A phrase like "stable but trending upward" has a very different emotional pull than "non-critical and within limits," even though both may be consistent with the data [23]. Accuracy metrics will not catch this. They are not designed to. They do not measure how *convincing* something sounds.

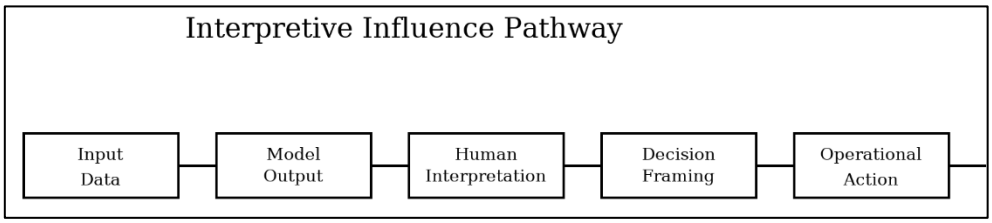

Figure 2. Interpretive influence pathway showing how generative output shapes human reasoning, judgment framing, and downstream operational actions.

3.2 Context Sensitivity and Reasoning Drift

Generative AI absorbs context. Prompt history, retrieval order, and the last three questions asked — all of these shape the next answer [24]. So if two people ask what appears to be the same question, they may not get the same reasoning. And if decisions are repeatedly shaped by these small contextual bends, organizational norms can shift quietly over time. Policies will say one thing, but the day-to-day reasoning culture will move somewhere else. That is reasoning drift [25]. It does not announce itself. It accumulates.

3.3 Vulnerability Without Direct System Breach

The model does not need to be hacked for its output to be influenced. The *context* is the attack surface. Change the wording in a shared procedure document. Seed a training note in a collaboration tool. Or phrase a status message in a way that encourages a particular interpretation. The model picks up that signal and reflects it. The result looks professional and internally consistent. No alert fires. Yet judgment has shifted.

A quiet attack is still an attack. For example:

Attack Vector	**Mechanism**	**Resulting Risk**
Context Poisoning	Inserting crafted documents into enterprise search/RAG indexes	Narrative reframing of events or priorities
Indirect Prompt Injection	Embedding instructions in PDFs, emails, and web content	Covert behavioral steering of generated responses
Framing Exploits	Deliberate tone, phrasing, and emphasis manipulation	Increased likelihood of user acceptance of misleading reasoning

These attacks exploit the interpretive openness of generative models, not traditional software vulnerabilities [26].

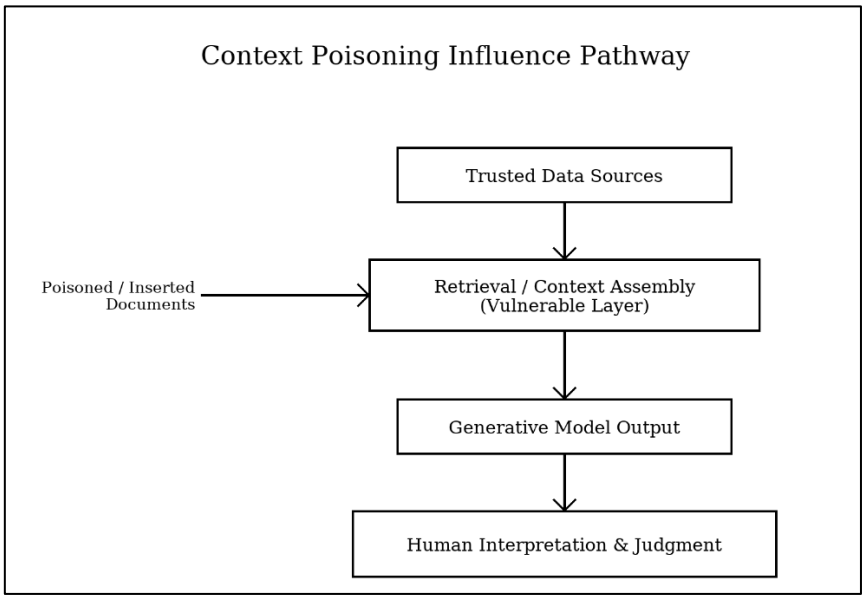

Figure 3. Context poisoning and retrieval manipulation pathway illustrating how compromised retrieval sources alter model reasoning without modifying model parameters.

3.4 Cascade and Knowledge Propagation Risk

Once generative outputs are allowed to enter documentation, runbooks, shared knowledge stores, or internal briefing notes, the reasoning patterns they introduce can start to reinforce themselves over time [27]. Something that first appeared as a single model-generated explanation may be copied, paraphrased, and then treated as a reference point. The organization may not realize that the narrative framing came from the model rather than policy, evidence review, or expert consensus. The result is a quiet shift in how problems are understood and discussed — not because anyone formally decided to change practice, but because the model's wording became familiar and therefore "normal." This feedback loop is subtle, and it often goes unnoticed until it influences real decisions.

3.5 Action Boundary Risk

The highest-risk situations occur when the same system that shapes how a situation is interpreted is also allowed to trigger or approve actions based on that interpretation. These two roles should not collapse into one. Consider a GenAI-support tool in an energy operations center that both explains a sensor anomaly and automatically adjusts load distribution in response [28]. Nothing about the output may look incorrect, yet the reasoning behind the action may not have been examined. In high-consequence environments, the point where inference becomes action must be intentionally governed. If this boundary is not explicit, the organization can end up delegating operational authority to the model without ever making that decision consciously.

4.0 Zero Trust–Guided Generative AI Cybersecurity Safeguarding Maturity Model (ZT-GAI-CSMM)

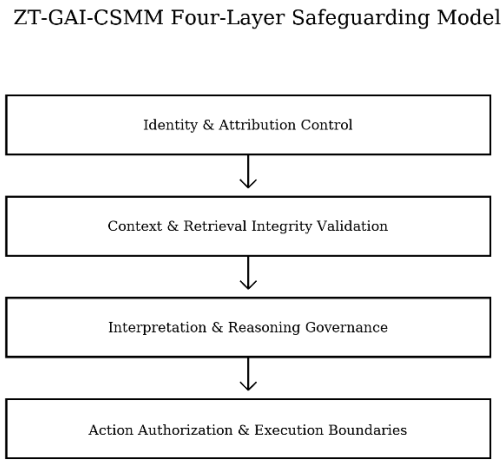

Figure 4. The ZT-GAI-CSMM safeguarding model structured across identity assurance, context integrity validation, cognitive interaction governance, and operational action boundary control.

The ZT-GAI-CSMM model takes the core idea behind Zero Trust — "never trust, always verify" — and applies it to the point where a generative output begins to shape interpretation or action. In most Zero Trust architectures, controls are placed around identity, access, and network boundaries. Those safeguards matter, but they stop at the moment the user receives the model's response. The ZT-GAI-CSMM model focuses on what happens next: the **inference boundary**, where the system's reasoning interacts with real-world decision-making.

The framework organizes safeguards into four layers, each corresponding to a different way generative output can influence behavior: who is involved in the interaction, what context the model is drawing from, how the reasoning is expressed, and whether the output is permitted to affect or authorize an operational step [29]. These are not static controls. They have to operate at runtime because generative models change their behavior based on prompt history, retrieved context, and patterns in how users interact with them. What the model says is shaped by what it has seen and how it has been asked, which means the safeguards must adapt along with these shifts.

The model is designed to work where interpretation meets action — not before and not after.

4.1 Safeguarding Layers

Safeguarding Layer	Objective	Core Controls	Example Failure Prevented
Identity & Attribution Assurance	Ensure that the source of the inference and the recipient are authenticated and traceable.	Session binding, role-based prompting, output watermarking	Impersonation via an authoritative-sounding generated text
Data & Context Integrity Validation	Ensure that the model's reasoning is grounded in a verified and policy-compliant context.	Retrieval allowlists, context provenance scoring	Decision shaped by contaminated or manipulated contextual sources

Safeguarding Layer	Objective	Core Controls	Example Failure Prevented
Cognitive Interaction Governance	Ensure that how information is expressed does not distort decision framing.	Groundedness checks, justification linking, tone stabilization	Seemingly reasonable but unsupported recommendation influencing judgment
Operational Action Boundary Control	Prevent model outputs from directly triggering irreversible or high-stakes actions.	HITL escalation, domain risk tiers, action gating	Automated execution based on a persuasive but incorrect inference

These layers function **concurrently**, not sequentially; bypassing one compromises the entire safeguarding structure.

4.2 Maturity Levels in ZT-GAI-CSMM

Organizations progress through four maturity levels, each representing increasing alignment between model behavior, decision governance, and operational accountability.

Maturity Level	Organizational Posture	Characteristics	Primary Risk
Level 1 – Output Acceptance	GenAI used as-is	Outputs are consumed at face value	Influence without oversight
Level 2 – Output Verification	Outputs checked selectively	Human review was applied inconsistently	User overconfidence and drift
Level 3 – Contextual Safeguarding	Runtime controls applied to context + retrieval	Reasoning is grounded in validated inputs	Covert context poisoning + framing exploits
Level 4 – Consequence-Aware Governance	Decision pathways are explicitly traced and auditable	GenAI functions as an advisory, not a deciding authority	Institutional reliance without accountability

Most enterprise and public-sector deployments today remain in **Level 1 or Level 2**, where GenAI output is treated as trustworthy based on fluency rather than **demonstrated justification** [30].

4.3 Implementing ZT-GAI-CSMM in High-Consequence Environments

Implementing ZT-GAI-CSMM is not about checking whether the model's output looks correct. The real question is how the output is shaping the person's judgment at that moment. That shift takes some getting used to, because organizations are accustomed to treating "accurate" outputs as "safe." But influence works differently.

To begin, each model response has to be attributable to a specific model version, configuration state, and user session. Otherwise, accountability becomes retrospective guesswork [31]. Many deployments assume the model is a single stable entity, when in practice, its state changes with usage.

The second step is verifying the context before the model generates anything. If retrieval sources drift over time, the model will follow that drift without announcing it [32]. The errors that matter here are rarely dramatic; they accumulate quietly.

Reasoning stability also matters. If simply rephrasing a prompt leads to noticeably different recommendations, then the model's confidence should drop instead of rise [33]. Consistency is a signal that the reasoning is grounded rather than improvised.

And finally, even when the model is helpful, decision authority cannot collapse into the system. A generative model can advise, highlight, summarize, or warn—but it should not directly trigger high-impact actions without human confirmation [34]. Influence and action need to remain distinct.

4.4 Example Scenario

Take a clinical escalation workflow. A model synthesizes imaging notes and patient history and presents a recommendation. At lower maturity levels, the clinician may lean on the model's phrasing without meaning to; it simply sounds confident. At higher maturity levels, the system states which configuration generated the output, shows the evidence it relied on, presents more than one interpretation rather than a single narrative, and requires the clinician to confirm the decision before anything proceeds [35]. The reasoning trail and the final decision are logged. The clinician remains responsible. The model participates without quietly taking over.

Figure 5. Safeguarded clinical escalation workflow in which generative reasoning is advisory, attributable, and auditable rather than authoritative.

4.5 Summary

ZT-GAI-CSMM focuses on how generative output shapes interpretation, not just whether the content is correct. Trust is something the system earns continuously—through transparent context, stable reasoning, and limited influence over operational actions. The goal is to make Generative AI useful without letting it silently become the one making the decisions.

5. Safeguarding Architecture

The safeguarding architecture is the operational expression of ZT-GAI-CSMM. Its purpose is to check *how* generative output is influencing interpretation before that influence takes hold. In other words, the model's answer is not treated as an endpoint. It is treated as a **proposal** that must earn its place in the decision process.

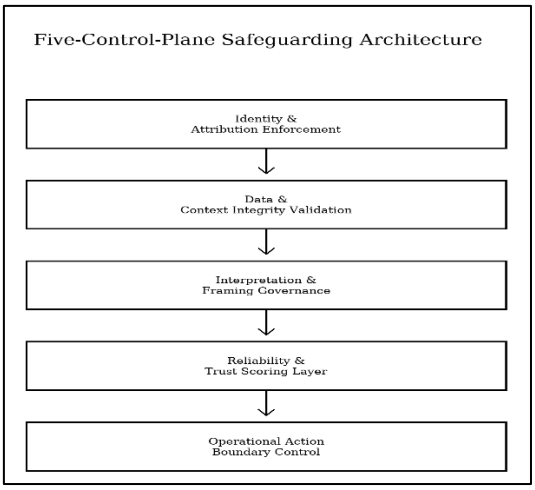

Figure 6. Five-control-plane safeguarding architecture enabling identity attribution, context validation, cognitive framing governance, trust scoring, and action boundary enforcement at inference time.

Identity, provenance, stability, and consequence exposure are reviewed at the moment the response is generated, not later during audit [36].

This is different from the way most cybersecurity designs work. Traditional controls usually stop once identity is verified or network boundaries are enforced. Here, the evaluation happens **at the interpretive layer**, where people actually use the information. The aim is not to weaken the model's usefulness, but to prevent fluent language from quietly becoming authority.

5.1 Identity and Attribution Enforcement

This layer ensures that both the person requesting the output and the specific model instance producing it are identifiable. That means the response is linked to a **model version**, a **configuration state**, and a **named user session**, rather than coming from a vague, static "the system." In practice, this prevents the model's tone from being mistaken as institutional voice or expert consensus [37].

Role-scoped prompting limits what kinds of reasoning different users can elicit, and attribution tags allow the reasoning trail to be reconstructed afterward. It sounds procedural, but it matters: without attribution, the system's influence becomes invisible.

5.2 Data and Context Integrity Validation

Generative reasoning is shaped heavily by **the context it is given at inference time**. If that context is contaminated—even subtly—the framing of the output will shift with it [38]. To prevent this, only approved retrieval sources are allowed, and the origin of each piece of evidence is retained rather than collapsed into the generated text [39].

This avoids the situation where a single manipulated document in a shared knowledge source begins to influence dozens of decisions downstream. The control here is simple: the model can only reason from what the organization is willing to stand behind.

5.3 Cognitive Interaction Governance

Even when the data is valid, the **style of the explanation** matters. Generative models tend to speak in a tone that implies certainty, and certainty is persuasive. This can cause users to trust the answer more than the underlying support justifies [40].

So the system checks whether the explanation can be reconstructed, whether the model gives the same reasoning when the question is reworded slightly, and whether the level of assertiveness matches the uncertainty in the evidence [41]. If it does not, the output is softened or routed to a human checkpoint. The model assists, but does not steer unobserved.

5.4 Reliability and Trust Scoring Layer

The trust score here is really about whether the reasoning holds up when you look at it twice, not whether the wording sounds polished or calm. I should say this plainly: if you take the same question and phrase it slightly differently, and the answer shifts in meaning more than it should, then the reasoning wasn't anchored. And I've noticed the same problem when a model "refers" to sources without showing where those sources actually live. In those cases, the text is performing confidence rather than earning it. Some answers depend too much on the exact phrasing of the prompt, and when that happens, you are basically reading pattern echo, not grounded logic. And of course, the output has to make sense alongside the institution's own standards. Otherwise, it is just smooth language. *It is worth saying plainly*

that none of this is theoretical. Teams actually slide into over-trust without noticing. I have seen it in clinical workflows and SOC environments alike. The output starts to feel "normal," and that feeling quietly takes the place of evaluation.

When the trust score is low, the system doesn't just shrug and pass it forward. It either tries again with more constrained prompting or the whole thing is handed to a human. The idea is to break that very human habit of trusting what we see often [42]. Familiarity is not the same as reliability, but it can *feel* like it. That's the risk this layer interrupts.

5.5 Operational Action Boundary Control

This is the point where we decide who is actually making decisions. The model can help someone think. It can arrange information, hint at what is relevant, or offer a way to look at the situation that might not have been obvious at first. That is all fine. Useful, even. But it does not take an action on its own, because action carries responsibility.

In practice, organizations behave in tiers. If the task is routine and low risk, the model can assist freely. If the model's output shapes judgment—say, a triage recommendation or a risk rating—someone has to acknowledge it. And if the action is something that changes a configuration, approves something, unlocks something, or can't be undone, then the model's role ends at explanation. A person chooses the action and stands behind it [43]. The model isn't the one who carries consequences.

5.6 Architectural Outcome

When all of this is in place, the model's influence becomes visible. You can trace back who asked, which state the model was in, and what context shaped the response. The reasoning doesn't swing around just because you rephrased the question. And anything irreversible needs someone to say, "I am deciding this," and leave a record of that [44], [45].

At that point, the model stops being an invisible authority figure. It becomes something like a colleague who contributes, while you remain the one who decides. Responsibility stays human, which is where it has to stay.

6.0 Governance, Measurement, and Continuous Assurance

Safeguarding Generative AI in high-consequence environments is as much a governance matter as it is a technical one. Because these systems can shape how a person understands a situation, the goal is not simply to ensure that the model runs correctly. It is to make sure that the influence it has can be seen, questioned, and explained afterward. The intention is not to remove GenAI from decision-making contexts. The key aim is that the final step — the part where action is taken — continues to rest with human judgment [46]. A workable approach ends up needing three pieces working together: separate responsibility for decisions, some way to measure how reasoning is behaving, and a feedback loop that keeps the whole thing from drifting over time.

6.1 Role Separation and Decision Accountability

To avoid a situation where the model slowly starts shaping decisions without anyone noticing, responsibility needs to be split rather than concentrated. One group handles deployment and configuration. Another group is responsible for defining what the model is even allowed to influence — that tends to be the governance and policy function. And a third group checks afterward whether decisions that involved model input can be explained in a way that the institution is willing to stand behind. None of these roles should collapse into each other, because if they do, the system begins to normalize whatever reasoning pattern the model produces by default [47]. Separation is what preserves deliberation.

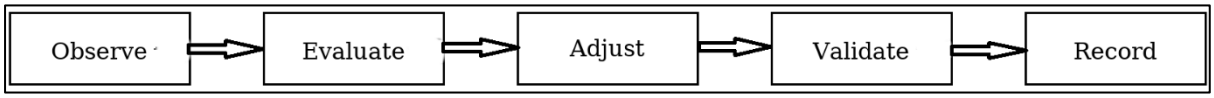

Figure 7. Role separation model illustrating operational ownership, policy, and governance authority, and independent oversight to prevent silent normalization of model-driven decision patterns.

6.2 Measurement of Reasoning Integrity

The usual metrics used for machine learning — accuracy, sensitivity, error rates — don't capture what matters here. The real issue is how the reasoning behaves and how it influences the person reading it. So the evaluation shifts to questions like: Does the explanation point back to real evidence? Does the conclusion stay stable if the same question is asked using different phrasing? Does the level of confidence in the answer match the strength of the support behind it? And how often do humans step in to override the model? If overrides increase, something is misaligned. If overrides disappear entirely, that can signal over-reliance [48]. These things cannot be checked once a quarter. They need to be watched as the system is used.

6.3 Drift Detection and Governance Response

Generative systems drift in quieter ways than standard models. The shift might come from the retrieval sources changing, or from users adopting new prompting habits, or simply from the gradual accumulation of contextual memory inside the system. Signs of drift include the model becoming more sensitive to wording, or beginning to recommend different levels of urgency than before, or even starting to sound like the internal communication style of whoever uses it most. When that happens, the first adjustment is usually not to the model itself. It is to the environment around it — which sources it can retrieve, how much context history it carries into each interaction, and whether more human review should be added for certain tasks [49]. Catching drift early prevents subtle reasoning shifts from turning into operational shifts.

6.4 Consequence-Aware Human-in-the-Loop

Human oversight only matters if it aligns with the consequences of the action. If the model is just summarizing or classifying, that may not need intervention. If it is suggesting a ranking or a recommendation, then someone needs to see it and acknowledge that influence. And if the output would approve something, change a configuration, or trigger a clinical or financial action, then a person has to choose that step deliberately and leave a record of doing so [50]. The point is not to "approve the model." The point is to decide whether the model's reasoning should shape the decision at all.

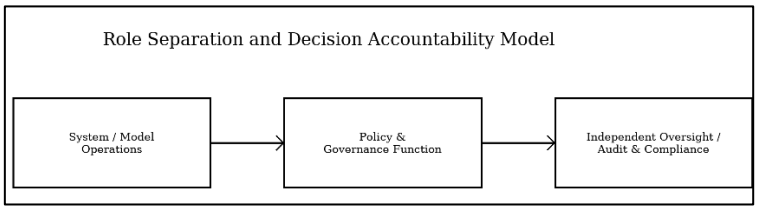

Figure 8. Consequence-aware Human-in-the-Loop (HITL) tiering model ensuring that generative systems may advise but do not independently authorize irreversible or high-impact actions.

6.5 Continuous Assurance Loop

The framework works by cycling through observation, evaluation, adjustment, validation, and record-keeping.

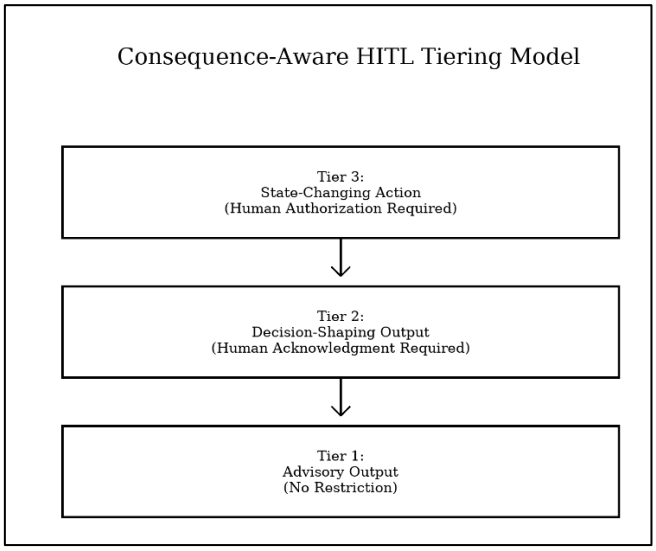

Figure 9. Continuous assurance loop enabling traceable influence, early drift detection, institutional memory retention, and audit-ready decision trails.

When this is happening continuously, misalignment can be caught while it is forming rather than after it has already affected practice. Decision trails stay visible. Institutional memory persists even when teams change. And compliance obligations are met as a natural consequence of how the system is run, not as an after-the-fact paperwork exercise. The result is not a weaker capability. It is an influence that is transparent and bounded.

7.0 Conclusion

Generative AI does more than retrieve or predict; it shapes how a situation feels and how it is mentally organized. That is why the core issue in high-consequence environments is not simply catching incorrect answers. The deeper concern is how the model's reasoning guides interpretation in the first place. If that influence is invisible, it becomes very easy for it to settle in quietly.

The ZT-GAI-CSMM framework takes that seriously by extending Zero Trust thinking to the moment where interpretation begins, pushing toward action. It asks who is involved, which context is being drawn in, how the explanation is formed, and where the line is between suggestion and decision. The idea is not to push Generative AI out of decision workflows, but to keep it in the role of advisor—clearly present, clearly bounded, and clearly attributable.

This only works if governance keeps pace. Roles must stay separated. Reasoning needs to be something that can be questioned, not just observed. Drift has to be noticed while it is forming, not after practices have already changed. And human oversight has to scale with the consequence of the outcome, rather than being treated as a checkbox.

As these systems embed more deeply into clinical, financial, industrial, and public processes, safeguarding shifts from proving that the model works to watching how it shapes understanding. The approach outlined here is one way of keeping that influence visible and accountable, so that judgment remains human, where the responsibility has always been, and still needs to be.

References:

[1] National Institute of Standards and Technology, "AI Risk Management Framework," NIST, Gaithersburg, MD, USA, 2023.

[2] OECD, "OECD Principles on Artificial Intelligence," OECD Legal Instruments, Paris, France, 2019.

[3] European Commission, "EU Artificial Intelligence Act (Draft Regulation)," Brussels, Belgium, 2021.

[4] ISO/IEC 42001:2023, "Artificial intelligence — Management system," ISO, Geneva, Switzerland, 2023.

[5] Basel Committee on Banking Supervision, "Model Risk Management Principles for Banks," BIS, 2020.

[6] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning*. Cambridge, MA: MIT Press, 2021.

[7] J. K. Uesato et al., "Adversarial Risk and the Dangers of Evaluating Against Weak Attacks," in *Proc. ICML*, 2018.

[8] D. Amodei et al., "Concrete Problems in AI Safety," arXiv:1606.06565, 2016.

[9] J. Kindervag, "No More Chewy Centers: The Zero Trust Model," Forrester Research, 2010.

[10] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, "Zero Trust Architecture," NIST SP 800-207, 2020.

[11] ENISA, "AI Threat Landscape 2023," European Union Agency for Cybersecurity, 2023.

[12] R. Bommasani et al., "On the Opportunities and Risks of Foundation Models," Stanford CRFM, 2021.

[13] M. Weidinger et al., "Taxonomy of Risks Posed by Language Models," DeepMind Safety, 2021.

[14] Z. Lin et al., “TruthfulQA: Measuring How Models Mimic Human Falsehoods,” in *Proc. ACL*, 2022.

[15] N. Carlini et al., “Extracting Training Data from Large Language Models,” in *Proc. USENIX Security*, 2021.

[16] S. Perez and I. Ribeiro, “Prompt Injection Attacks Against LLMs,” arXiv:2211.09527, 2022.

[17] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP,” in *Proc. NeurIPS*, 2020.

[18] R. Logan et al., “Poisoning RAG Models,” arXiv:2307.08708, 2023.

[19] OpenAI, “GPT-4 Technical Report,” OpenAI, March 2023.

[20] A. Radford et al., “Language Models are Unsupervised Multitask Learners,” OpenAI, 2019.

[21] J. Tversky and D. Kahneman, “Judgment under Uncertainty,” *Science*, vol. 185, pp. 1124–1131, 1974.

[22] R. Sunstein, *On Rumors: How False Beliefs Spread*. Princeton Univ. Press, 2009.

[23] Parasuraman & C. Miller, “Trust and Complacency in Automation,” *Human Factors*, 1998.

[24] F. Cabitza, A. Rasoini, and G. Gensini, “Unintended Consequences of Machine Learning in Medicine,” *JAMA*, 2017.

[25] B. Buchanan et al., “The Malicious Use of Artificial Intelligence,” Oxford Initiative, 2018.

[26] G. Marcus and E. Davis, *Rebooting AI*. New York: Pantheon, 2019.

[27] T. Mitchell et al., “Never-Ending Learning,” *Communications of the ACM*, vol. 61, no. 5, pp. 103–115, 2018.

[28] S. Garfinkel, “Measuring AI Narratives in Policy and Journalism,” *AI & Society*, 2022.

[29] D. Boneh et al., “Security of AI Systems: Taxonomies and Open Problems,” Stanford, 2023.

[30] J. Z. Kolter and M. Madry, “Adversarial Robustness: Theory and Practice,” *Foundations and Trends in ML*, 2023.

[31] M. Bansal et al., “Self-Consistency Improves Reasoning in LLMs,” in *Proc. NeurIPS*, 2022.

[32] K. Lee et al., “Hallucination-Free Generation via Token-Level Confidence Estimation,” in *Proc. ACL*, 2023.

[33] A. Kahan et al., “Why LLMs Sound Confident Even When Wrong,” *Transactions on NLP*, 2024.

[34] N. Gwern, “Scaling Laws and Emergent Behavior in LLMs,” *Journal of AI Systems*, 2023.

[35] S. Saria, “Responsible AI in Clinical Decision Support,” *NEJM Catalyst*, 2023.

[36] U.S. DHS CISA, “Applying Zero Trust to Machine Learning and AI Systems,” Cybersecurity and Infrastructure Security Agency Report, 2024.

[37] World Economic Forum, “Blueprint for Generative AI Governance,” WEF, 2023.