CSAI Foundation

Cloud Security Alliance AI Safety Initiative

Salami Slicing: Cumulative Trust Exploitation in LLMs

How Incremental Low-Risk Prompts Defeat Single-Turn Safety Guardrails

Unofficial AI-assisted Research

2026-04-15

Key Takeaways

- A newly named attack class—"salami slicing"—exploits the fundamental mismatch between LLMs' stateful multi-turn context accumulation and their predominantly single-turn safety evaluation architecture.
- Research published in April 2026 demonstrates attack success rates exceeding 90% against frontier models including GPT-4o, Gemini 2.5 Pro, DeepSeek V3, and Qwen 3 using this technique [1].
- The attack operates by chaining individually benign prompts that each stay below refusal thresholds; only when evaluated as a cumulative sequence do they exceed safety boundaries—a property that standard per-turn guardrails cannot detect.
- A proposed defense, Cumulative Query Auditing (CQA), audits the concatenated history of all prior prompts at each turn; early experiments reduce attack effectiveness by at least 44.8% [1].
- Agentic AI deployments—with their long-running sessions, rich context windows, and autonomous tool use—face disproportionate exposure to this attack class.
- Organizations should treat conversation-level monitoring as a first-class security requirement, not an afterthought, particularly for any LLM deployment that supports extended multi-turn sessions.

Background

The term "salami slicing" originates in financial fraud, where small, individually imperceptible transactions are aggregated over time to yield material theft without triggering any individual detection threshold. The structural analogy to adversarial prompt engineering is precise: both exploit the gap between what a detection system evaluates in isolation and what actually accumulates across a sequence of actions.

Large language models are trained and evaluated predominantly on single-turn interactions. Safety fine-tuning, RLHF reward modeling, and most deployed content filters assess each prompt or response in isolation, checking whether the immediate input or output crosses a harm threshold. This architecture reflects the dominant deployment pattern during the training era: a user submits a question, the model

answers, and the session is treated as essentially stateless from a safety perspective. As LLMs have migrated into long-running assistant deployments, agentic pipelines, and multi-turn chat applications, this assumption has become a structural vulnerability.

Multi-turn jailbreak attacks as a class have been studied since at least 2024, with techniques ranging from gradual role-play escalation to the "Echo Chamber" approach—planting benign but semantically loaded fragments early in a conversation that are later referenced to amplify harmful content [2]—and the "Bad Likert Judge" technique, which leverages a model's own evaluation capabilities across successive turns to progressively elicit unsafe outputs [11]. These approaches share a common dependency: they require some degree of explicit harmful signal in the final turn or rely on carefully engineered contextual scaffolding. The salami slicing framework, as characterized in recent research, represents a qualitative advance because it removes this dependency entirely.

The fundamental vulnerability being exploited is what researchers describe as the recency bias in LLM safety evaluation: models are primarily trained to recognize harmful content within the immediate context of a single exchange, leaving them unable to reason about whether a sequence of individually acceptable requests has cumulatively constructed a harmful outcome [1]. This is not a quirk of any particular model—it reflects how safety training datasets are structured and how inference-time content filters are architecturally positioned.

Security Analysis

The Attack Mechanism

The salami slicing attack, as formalized in arxiv preprint 2604.11309, operates through four stages [1]. First, an attacker constructs an initial safe prompt that establishes a harmless, topic-aligned foundation—something innocuous enough to establish the conversational frame without raising refusal flags. Second, the attacker generates a sequence of k incremental perturbations: small requests that each nudge the model's behavior incrementally toward the attacker's ultimate objective while remaining, in isolation, below refusal thresholds. Third, each perturbation is repeated t times in the sequence to reinforce the accumulated context without introducing any single large escalation step. Finally, the complete sequence is delivered to the target model, which has by that point been progressively conditioned to comply with the final, harmful request.

The attack is described as "trigger-free and context-agnostic" because it does not require a single explicit harmful prompt, nor does it depend on brittle model-specific jailbreak scaffolding. The accumulation itself is the mechanism. Researchers report attack success rates of 86.9% against GPT-4o

on the AdvBench benchmark under five-shot conditions—consistent with the paper's aggregate figure of over 90% against GPT-4o across all evaluated scenarios [1]—and 91.2% against Gemini 2.5 Pro, 95.2% against Qwen 3, and 96.0% against DeepSeek V3 under five-shot conditions. These figures represent material advances over single-turn approaches and suggest broad applicability across the frontier model landscape.

Distinction from Classic Prompt Injection

Classical direct prompt injection and single-turn jailbreaks are fundamentally attempts to conceal or reframe a malicious request within a single interaction. The attacker must embed sufficient harmful signal in one turn to overcome alignment constraints, which is why elaborate jailbreak templates—role-play scenarios, hypothetical framings, token obfuscation—have historically been necessary. Defenses against single-turn attacks can therefore be anchored to per-turn harm classification: if the current input scores above a harm threshold, refuse.

Salami slicing inverts this logic. No individual turn contains sufficient harmful signal to trigger per-turn defenses. The attack exploits the model's tendency toward contextual consistency—its learned propensity to treat earlier conversational context as authoritative framing for later requests—to progressively condition compliant behavior. The closest established analogue in the broader multi-turn jailbreak literature is the "context fusion" approach, which masks malicious intent through keyword substitution and scenario-based context spread across multiple turns [3]. Salami slicing is more general: it requires no keyword substitution or scenario engineering, relying instead on cumulative behavioral conditioning.

A related but distinct phenomenon—"contextual contamination"—has been observed in settings where LLMs ingest high-density datasets, causing the model's attention mechanisms to prioritize the statistical patterns of the input context over static system instructions [4]. While that phenomenon arises in data ingestion rather than adversarial interaction, it points to the same underlying architectural property: LLM safety alignment is anchored to training-time distributions, not to runtime reasoning about cumulative conversational state.

Implications for Agentic AI

The security implications of salami slicing are substantially amplified in agentic AI deployments. Autonomous agents by design engage in long-running, multi-step task execution where each step builds on prior context. Session lengths that would be unusual in a chat application are routine in agents

managing workflows, executing code, browsing the web, or interacting with enterprise systems. The longer the session, the more accumulated context the attacker can leverage, and the harder it becomes for any per-turn guardrail to distinguish legitimate task decomposition from adversarial conditioning.

Agentic systems also introduce delegation chains and trust inheritance. When a primary agent delegates subtasks to specialized sub-agents, the delegating context—including any adversarially accumulated framing from earlier in the session—may be propagated as authoritative instruction. This creates pathways for salami slicing attacks to traverse trust boundaries between agents, using accumulated context in one agent's session to condition behavior in another's. The CSA MAESTRO framework characterizes these as "cross-layer threats"—chains of exploitation where a vulnerability at one layer enables compromise at another—and notes that LLM-level context manipulation is a primary entry point for this class of attack [5].

Detection is further complicated by the lack of clear ground truth for "accumulated harm" in real-world deployments. A user legitimately debugging a complex technical problem may send dozens of individually benign requests that, viewed in sequence, look superficially similar to an incremental escalation. The signal-to-noise ratio for anomaly detection at the session level is much lower than at the turn level, requiring more sophisticated instrumentation to maintain acceptable false-positive rates.

The CQA Defense and Its Limits

The Cumulative Query Auditing (CQA) defense proposed alongside the salami slicing characterization takes the most direct approach: at each conversation turn, concatenate all prior prompts and evaluate the full cumulative sequence against a harm classifier, refusing if the aggregate score exceeds a threshold [1]. This directly addresses the recency bias vulnerability by forcing evaluation over the complete interaction history rather than just the current input. Experiments show CQA reduces the salami attack's effectiveness by at least 44.8%, with a maximum blocking rate of 64.8% against other multi-turn jailbreak techniques.

Two practical limitations are significant. First, computational cost scales with session length: evaluating a growing concatenated history at each turn imposes latency and token cost overhead that may be prohibitive in high-throughput production deployments. Second, CQA's blocking rate—while meaningful—leaves a residual attack surface. At 64.8% maximum blocking, roughly one in three sophisticated multi-turn attacks in the evaluated conditions still succeeds. This underscores that CQA is a mitigation, not a solution, and must be layered with other defenses to approach adequate protection.

More fundamentally, the salami slicing finding surfaces a design debt in current LLM safety architecture: the industry has invested heavily in per-turn alignment and refusal training, but relatively little in session-level behavioral monitoring frameworks. Closing this gap requires changes at the evaluation, training, and

deployment layers.

Recommendations

Immediate Actions

Organizations operating LLM deployments that support multi-turn sessions should audit current guardrail architectures to determine whether content filtering operates exclusively at the turn level. If so, the deployment is exposed to the salami slicing class of attacks without any detective control. Enabling conversation-level logging—capturing the full interaction history, not merely the current turn—is a prerequisite for any subsequent session-aware defense and should be treated as an immediate baseline control.

For high-risk deployments such as those with access to privileged systems, sensitive data, or the ability to execute code or transactions, security teams should conduct targeted red-team exercises explicitly testing cumulative escalation patterns. Standard single-turn jailbreak red-teaming does not adequately surface this risk. Testing should include both manually crafted escalation sequences and automated tools that implement the staged perturbation approach described in the research literature.

Short-Term Mitigations

Deploying a version of Cumulative Query Auditing—even a lightweight implementation using an existing harm classifier applied to the concatenated session history—provides measurable risk reduction with relatively straightforward engineering. The latency tradeoff can be managed by batching full-history evaluations at configurable intervals (e.g., every five turns) rather than every turn, accepting that the gap introduces a window of exposure while managing cost.

Session length limits constitute a practical architectural control. Restricting the number of turns or the total token volume per session before requiring a context reset reduces the attack surface available for incremental conditioning. This is especially applicable to agentic deployments: implementing checkpoints that require explicit human confirmation before an agent continues past a defined session depth adds both a detection opportunity and a friction layer that raises the cost for attackers. Organizations should also consider deploying separate session-context classifiers—models specifically trained or fine-tuned to recognize escalation patterns in conversation trajectories—as a dedicated detection layer distinct from the primary model's own safety evaluation.

System prompt hardening can add friction without eliminating risk. Explicit instructions to the model to maintain awareness of cumulative conversational direction, to treat repeated incremental requests for sensitive information as suspicious, and to invoke clarification or refusal when a session appears to be progressively escalating toward a sensitive domain may reduce susceptibility at the model layer, though this is not a reliable sole control given the attack's demonstrated effectiveness against models with robust alignment.

Strategic Considerations

The salami slicing finding is a signal that the industry's current safety evaluation paradigm—predominantly single-turn, predominantly static—requires structural extension. Procurement and vendor assessment processes should include explicit questions about whether safety evaluation is applied at the session level, whether session interaction histories are logged and analyzed, and whether vendors have tested their deployments against multi-turn escalation scenarios. The OWASP Top 10 for LLM Applications 2025 [10] lists Prompt Injection as the top risk; the corresponding LLM01:2025 risk entry [6] covers prompt injection broadly, though multi-turn cumulative escalation represents an attack surface that extends beyond the single-turn scenarios most commonly addressed in vendor compliance frameworks. Compliance claims that address only the most straightforward single-turn injection scenarios should therefore be treated as incomplete.

For agentic system architects, session isolation between agents—limiting the context that any one agent can inherit from another—is a critical architectural pattern. Trust should not propagate automatically through delegation chains without explicit re-verification at each handoff. This principle is consistent with Zero Trust governance models for AI agents and reduces the blast radius of any single-session compromise by limiting the accumulated context that can be weaponized across the agentic pipeline [7].

At the governance layer, organizations should incorporate session-level behavioral monitoring into their AI risk management programs. This includes defining what "normal" multi-turn interaction patterns look like for a given deployment, establishing anomaly detection baselines, and creating escalation procedures when session-level harm classifiers trigger. These monitoring investments serve double duty: they address salami slicing specifically while also improving visibility into other emerging multi-turn attack patterns such as Echo Chamber manipulation and context fusion attacks.

CSA Resource Alignment

This research note connects to several active CSA frameworks and publications.

The **MAESTRO Agentic AI Threat Modeling Framework** provides the most direct structural context for the risks described here [5], with applied guidance on mapping MAESTRO threat scenarios to real-world agentic architectures available through CSA's extended resources [12]. MAESTRO's characterization of cross-layer threats—where exploitation chains move fluidly between the Foundation Model layer (where salami slicing operates) and the Agent Orchestration and Tool Integration layers—maps directly onto the amplified risk in agentic deployments. Organizations applying MAESTRO threat modeling to their AI systems should add cumulative context manipulation as an explicit threat scenario in the Foundation Model and Agent Memory layers.

The **AI Controls Matrix (AICM)**, with its 18 domains and 243 control objectives covering the full AI supply chain, addresses prompt injection and session management controls that are directly applicable to salami slicing defenses [8]. The AICM's Shared Security Responsibility Model provides a useful lens for allocating responsibility for session-level monitoring: model providers own alignment training, orchestration-layer providers own session management controls, and application providers own the deployment-specific guardrail configuration. No single party can fully close the risk unilaterally, making clear ownership allocation essential.

The **CSA Agentic Trust Framework** and its Zero Trust principles for AI agents provide architectural guidance on limiting context propagation across delegation chains—directly addressing the cross-agent amplification risk identified in this note [7]. The principle of least-privilege context inheritance, analogous to least-privilege access control in traditional Zero Trust architectures, should be a standard design pattern for multi-agent systems exposed to adversarial environments.

The **STAR for AI** program provides an assurance pathway for organizations seeking independent verification of their AI security posture [9]. Given the session-level monitoring gaps this note identifies, STAR for AI assessments should explicitly evaluate whether assessed organizations have instrumented and tested multi-turn attack resilience, not solely single-turn prompt injection defenses.

Finally, CSA's **AI Organizational Responsibilities** publications—covering governance, risk management, compliance, and cultural aspects—provide the board and executive framing for the investments in session-level monitoring infrastructure that addressing this risk requires. The case for these investments is strengthened by the quantitative data now available: attack success rates above 90% against frontier models using a technique that currently evades most deployed defenses represent a material, documentable risk that belongs in enterprise AI risk registers.

References

[1] Zhang, Y., Wang, K., Wu, J., Wu, H., Zhou, Y., Wei, Z., Wu, D., Chen, X., Sun, J., and Sun, M. "The Salami Slicing Threat: Exploiting Cumulative Risks in LLM Systems." arXiv preprint arXiv:2604.11309, April 2026.

[2] Alobaid, A., Jordà Roca, M., Castillo, C., and Vendrell, J. "The Echo Chamber Multi-Turn LLM Jailbreak." arXiv preprint arXiv:2601.05742, January 2026.

[3] Sun, X., Zhang, D., Yang, D., Zou, Q., and Li, H. "Multi-Turn Context Jailbreak Attack on Large Language Models From First Principles." arXiv preprint arXiv:2408.04686, August 2024.

[4] Jacoby, K. "Contextual Contamination: The Silent Drift of Large Language Models via Stored Conversation Data." PhilArchive preprint, accessed April 2026.

[5] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 2025.

[6] OWASP. "LLM01:2025 Prompt Injection." OWASP Gen AI Security Project, 2025.

[7] Cloud Security Alliance. "The Agentic Trust Framework: Zero Trust Governance for AI Agents." CSA Blog, February 2026.

[8] Cloud Security Alliance. "AI Controls Matrix." CSA, 2025.

[9] Cloud Security Alliance. "CSA STAR for AI." CSA, 2025.

[10] OWASP. "OWASP Top 10 for Large Language Model Applications." OWASP Foundation, 2025.

[11] Palo Alto Networks Unit 42. "Bad Likert Judge: A Novel Multi-Turn Technique to Jailbreak LLMs by Misusing Their Evaluation Capability." Unit 42 Research, 2025.

[12] Cloud Security Alliance. "Applying MAESTRO to Real-World Agentic AI Threat Models." CSA Blog, February 2026.

