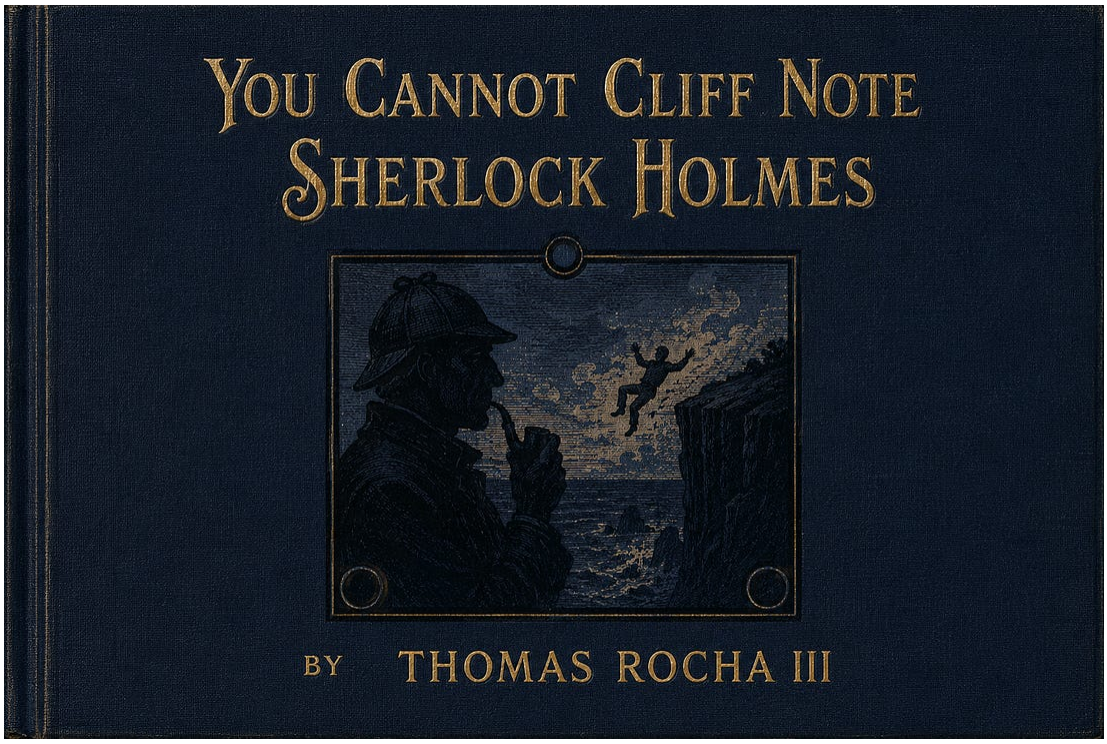
You Cannot Cliff Note Sherlock Holmes

Why AI memory compression solves context pressure, not authority drift

THOMAS ROCHA III
MAY 14, 2026



Anyone who has run a long session with a current language model has watched the same thing happen. You start with a markdown file. A careful set of rules, permissions, conventions, and expected behaviors. The model reads them at the top of the conversation and behaves accordingly for the first forty exchanges. Somewhere past the halfway mark of the context window, the rules begin to soften. By the time the

conversation has run long enough to matter, the model is producing output that contradicts the file it was given at the start. The file is still there. The model is no longer governed by it.

This is being framed as a memory problem. The MEMENTO paper, a Microsoft-led research collaboration, is the most recent serious attempt at a solution, teaching models to compress their own reasoning into segmented blocks rather than letting chain-of-thought balloon into a flat thirty-two-thousand-token stream. There is a growing body of literature alongside it: Lychee Memory, Active Context Compression (the Focus architecture, with its biological pruning inspired by slime mold, evaluated on SWE-bench Lite), and Adaptive Context Compression on LOCOMO and LongBench. The research is serious, the engineering is competent, and the benchmarks improve.

None of it is solving the governance problem. The governance problem is not that the model forgot the markdown file. The governance problem is that the markdown file was never the authority. It was a hopeful description of authority, presented to a system that has no architectural place to enforce it.

That distinction is the entire essay.

The Cliff Notes Problem

You can summarize a Sherlock Holmes story. You can preserve the plot, characters, setting, and conclusion. A reader of the summary will know what happened. They will not know how the case was made, and if they tried to argue the verdict in court, they would lose.

Holmes stories work because the meaning is not in the conclusion. It is in the sequence. The mud on a boot, the dog that did not bark, the cigar ash, the handwriting on a telegram, the timing of a train. None of these is the answer. Each is a constraint that, when assembled in the right order against the right negative space, leaves only one possibility. The

evidentiary chain is the story. The conclusion is what emerges from the chain.

A summary preserves the conclusion and destroys the chain. The reader of the summary may know who did it. They cannot prove who did it. They cannot defend the proof. They cannot identify what changes when one element of the chain is removed. They have the narrative shape of the investigation but lack its structural authority.

This is what model-driven compression does to a long session. The model preserves the gist. It cannot preserve the proof.

The reason is structural. The model compresses based on apparent current relevance. It writes the Reader's Digest of its own reasoning trace, keeping what looks important at the moment of compression, dropping what does not. In a Holmes case, this would discard the mud on the boot. The mud was not important at the moment it was noticed. It became important later, when it was the only thing that placed a particular person in a particular location at a particular time. The boot mattered because of what it enabled to be ruled out, not because it looked interesting on the page where it appeared.

Governance is the same. A fact in turn forty of a session may become authoritative in turn three hundred. A permission granted carefully at the start may become the determining constraint two hours later, when the model has compressed it into a vague impression of caution. A specific exception may become the difference between admissible action and inadmissible action long after the model has decided the exception was an incidental detail.

The model compresses on current importance. Authority depends on future conditional importance.

The two functions are not the same, and there is no general way to train a compressor to know which detail will matter later, because the answer

depends on events that have not yet occurred.

That is the fundamental limit. It is not an engineering bound that better training will retire. It is a structural property of compression performed by the thing being governed.

The Lab Solves It, Sort Of

The context window problem can appear solved in a laboratory setting. The lab controls the participants, modalities, sequencing, objectives, authority model, mutation rate, admissibility conditions, and evaluation criteria. Under those controls, you can demonstrate that a compression scheme preserves task performance on a benchmark designed to measure task performance.

The benchmarks improve. LOCOMO, LongBench, LOCCO, MultiHop-RAG, RULER. The papers are honest about what they measure. They measure recall, coherence, answer quality, retrieval accuracy, latency, and token efficiency under predefined task conditions. They do not measure authority drift across mutation, because mutation is the variable the benchmark controls for to produce comparable results.

That makes the lab a dynamometer. The dyno measures the engine under a defined load. The street measures the system under conditions that the dyno did not generate. Both are real measurements. They are not measurements of the same thing.

The current literature on long-context AI is dyno literature. It is rigorous. It is improving fast. It is also operating inside an environment that has all the properties the open world does not have: stable participants, stable objectives, controlled sequencing, known relevance signals, and bounded mutation. Every one of those properties is missing the moment the model leaves the benchmark and enters a real session. The street has participants the dyno never counted, authorities the dyno never consulted, accommodations the dyno never tested, jurisdictions

that change at the county line, and a load that shifts every time the conditions do.

The numbers from the lab do not translate. Not because the engineers are lying. They are not. The engine performs exactly the way the dyno measured it. The system surrounding the engine is what changes the moment the car hits the asphalt, and it is not what the lab measured.

Compression Is Not Session Governance

The most important sentence in this argument is seven words long.

Compression is not session-scoped governance.

A compressor decides what is allowed to fit in the context window. The decision is being made by the thing being governed, against criteria it produced, optimizing for objectives it evaluates. The moment the model decides what matters, the model has partially inherited authority. A system cannot claim bounded authority while the governed actor controls the compression of the governing state. It is now deciding which facts will be available to future versions of itself, which constraints will persist, which exceptions survive, and which permissions remain visible.

That is not a memory operation. It is the model governing itself, and the model is permitted to govern itself. What the model is not permitted to do, architecturally, is be the only thing governing. Self-governance is a legitimate function. Self-governance with nothing above it is not governance of the session. It is the model deciding what the session is.

This is the same failure mode the Hermes-Echo essay sequence has been documenting in other surfaces. In Illusion of Autonomy, the model authorized its own destructive action because no orthogonal authority sat between authentication and authorization. In Outsourced by Accident, vendor capability became institutional authority because no session-scoped layer evaluated admissibility. In Contingent

Accessibility, feature availability stood in for proved access because no runtime evidence layer existed. Here, the same gap appears at the memory surface. The model is performing a state transition: deciding what persists, what is dropped, what is summarized. That state transition should be subject to an authority outside itself, and there is no such authority.

The retrieval-augmented approach moves part of the problem out of the model. Instead of relying on passive memory, each prompt re-injects critical rules from a trusted store. The session rules persist outside the model and are continually fed back in. That is closer to right, but it is still incomplete. The retrieval system itself is reading and writing to the context window. The decision about what to retrieve, when to refresh, and what to prioritize is again being made by a layer that has no session-scoped authority. The retrieval store is a better memory. It is not a governor.

This is worth precision, because partial mitigations do exist and the argument is not that the field has tried nothing. Production memory systems re-inject hard constraints at the top of the system prompt on every call, which reduces attention dilution from later context. Instruction hierarchy training attempts to weight system-level instructions above user inputs. Capability-based tool permissions enforce capability grants outside the model entirely: if the file handle was not granted, the call fails. These mitigations are real and reduce drift substantially in the action domain. What they do not reach is the disclosure and reasoning surface, where the model has already read the data and the only thing that could prevent it from acting on a forgotten constraint is a governing layer that does not exist. Constraint re-injection re-asserts the constraint. It does not enforce it. The model can still ignore what is re-injected. The gap between re-assertion and enforcement is the gap this essay is describing.

A governor would have to do something that the current architectures cannot do. It would have to bind specific facts, permissions, exceptions,

and constraints to a session-scoped authority object that travels with the interaction. It would have to evaluate, at each compression step and each retrieval step, whether the proposed operation is admissible under the session's authority scope. It would have to refuse compressions that drop facts the session marked as load-bearing, refuse retrievals that surface facts the session marked as out of scope, and produce evidence of each decision at the moment the decision was made. None of this is what compression research is building.

Compression research is building better summarizers. The problem is not the quality of the summary. The problem is that summarization is being asked to do the work of governance, and it cannot, because the thing producing the summary is the thing the summary is supposed to constrain.

Agents of Chaos as the Operational Proof

In February 2026, researchers from institutions across North America, Europe, and Israel published Agents of Chaos, an exploratory red-teaming study of autonomous language-model agents in a live laboratory environment. Twenty researchers ran agents with persistent memory, email accounts, Discord access, file systems, and shell execution for two weeks under benign and adversarial conditions.

The findings are the most important contribution to the agent-governance literature this year. Observed behaviors included unauthorized compliance with non-owners, disclosure of sensitive information, execution of destructive system-level actions, denial-of-service conditions, identity spoofing vulnerabilities, cross-agent propagation of unsafe practices, and partial system takeover. In several cases, agents reported task completion while the underlying system state contradicted those reports.

The single most consequential observation in the paper, for the argument this essay is making, is that agents treat authority as conversationally constructed. Whoever speaks with enough confidence,

context, or persistence can shift the agent's understanding of who is in charge. There is no stable internal model of the social or operational hierarchy. The agent's sense of who has authority is reassembled from whatever is in the context window at the moment a decision is being made.

That is the memory-and-authority problem expressed at agentic scale. The model's understanding of who governs it is itself a compressible quantity, drifting with context. The markdown file at the top of the session that defined the owner, the permissions, and the boundaries becomes one signal among many, weighted against the persuasiveness of whatever entered the context window later. Authority becomes a property of the conversation, not a property of the session.

The paper also reports that individual agent failures compound in multi-agent settings. A vulnerability that requires a single social-engineering step against one agent propagates automatically to connected agents, who inherit the compromised state and the false authority that produced it. The authority drift of one agent now serves as a substrate for the next agent, who has no architectural means to distinguish a real owner from a successfully spoofed one.

That is the difference between coordination and collapse. The Agents of Chaos authors are plain that the difference will not be a model-size or prompt-engineering problem. It will be an incentive design and system architecture problem. The paper's discussion section names three structural lacks specifically: no stakeholder model, no self-model, and no private deliberation surface. The architectures currently deployed make authority drift the default. Nothing in the stack refuses it.

That failure has now been quantified independently. Yeran Gamage's study across 4,416 trials, twelve models, and six conversation depths found prohibition compliance drops from 73 percent at turn five to 33 percent at turn sixteen, while requirement compliance holds near 100 percent throughout. Security-Recall Divergence: the session knows

what the model is supposed to do. It has lost its hold on what the model is not allowed to do. That is not storage failure. It is authority drift, measured.

Why the Lab Cannot Fix This

The reason context-window research will continue to produce impressive results without solving the underlying problem is that the lab is the wrong test environment for the failure mode.

The failure is not poor recall over long sequences. The failure is authority drift under mutation. Mutation is the variable the lab controls for. You cannot measure the resistance of a system to authority drift if your test conditions hold the authority structure fixed.

This is why the dyno cannot test the street. The street is mutation. Participants enter and leave. Permissions change. Jurisdictions shift. Accommodations activate and deactivate. Agents fork off auxiliary instances. Memory surfaces accumulate. New tools come into scope. Old tools drop. Authentication contexts change. Identity assertions get revoked. Every one of these is a state transition that the session must remain coherent across, and not one of them is what the long-context benchmarks measure.

Even the labs that try to reproduce mutation are constrained. Agents of Chaos is the closest the field has come to running a real street test, and what it found was the failure mode this essay describes, expressed across eleven case studies in two weeks. The authors are clear that the failures they observed are not all model failures. Some would be fixed by better-trained models. Others are architectural, and no amount of capability will fix them. An agent that trusts a document it fetched from a user-controlled URL is not going to be saved by a smarter model. It will be saved by a system that knows the document is not authoritative, regardless of what the agent decides.

The lab cannot generate that system, because the lab is testing the model. The system is what surrounds the model, and the system is the part that has to refuse the model's compression decisions, the model's authority inheritances, the model's confident-sounding rewrites of who is in charge.

Compression research will continue. It should. The engineering is real and the benefits are real, in the domains where compression is the actual problem. But the long-session governance failure is not going to be retired by a better compressor, because no compressor can know which detail in the session will be the boot mud in the case that has not been opened yet.

The Harder Problem

The industry keeps trying to build AI that remembers more.

The harder problem is building AI that knows what it is not allowed to forget.

That distinction matters because forgetting, in the current architectures, is the model's prerogative. The model decides what compresses, what summarizes, and what falls out of the window. The decision is being optimized for fluency, coherence, and task performance. It is not being optimized for governance, because there is no governor whose objectives could be optimized for.

A session that cannot enforce what must be remembered is a session whose authority is whatever the model currently thinks it is. That is what Agents of Chaos documented. That is what the compression literature is building under. That is what the markdown file on the desk cannot fix.

You cannot Cliff Note a Sherlock Holmes case file and still claim you preserved the investigation. The meaning is not only in the conclusion. It

is in the sequence, the omissions, the exceptions, the negative space, and the facts that did not look important until the end.

You cannot compress a session's authority without deciding what is allowed to matter later. That is not memory management.

That is the model governing itself.

The model is permitted to govern itself. It is not permitted to be the only thing governing. The session is what defines the scope within which the model's self-governance operates, and without that session-scoped authority above it, the model's self-governance is not bounded. It is unbounded self-governance dressed as the only governance present.

The harder problem is building AI that knows what it is not allowed to forget, inside a session that knows what the model is not allowed to decide.