VELOCITY EDGE

CONTEXT ENGINEERING, PART 2

Why Do Long-Running Agents Drift?

ABSTRACT

The studies this part cites keep an instruction or task fixed and change the conditions around it: context length, where needed evidence sits, and the number of turns. Behaviour moved with length. Not every model or task showed this. Over turns the evidence was confounded. Evidence on position was mixed. Outside recent work on compaction, many cited studies measure task accuracy, not whether a rule was kept.

The part states a hypothesis about a rule the model must keep by reading it, and whose breaches can be observed, within a model's stated context length. On the hypothesis, such a rule would become more likely to be broken as the context lengthens. It would also do so as the rule falls further behind the model's proposal. The hypothesis makes no distance claim for a rule given once at the start. Each compaction since the rule was last present word for word would raise the risk, other things equal.

Some evidence runs against the hypothesis. On two models, one rule-keeping test attributed most of a fall to content, not length. Where a model judge found a summary carried the rule, one preprint found no violations. Its sample was small. The hypothesis is unfitted and untested. It says what would show it wrong.

Long work can run as a chain of short contexts built from a written record. One account shows it in use. One confounded measurement bears on its effect. Where a context must stay alive, the part argues for summarising last.

Phillip Pienaar
Velocity Edge

2026-11-19
Version 1.0.1

Contents

Executive Summary 3
1. The Question 3
2. What the Evidence Shows 4
3. Failures and the Settings They Belong To 5
4. A Hypothesis of Decay 6
5. A Chain of Short Contexts 8
6. Compaction as a Last Resort 9
7. What a Chain Does Not Fix 9
Conclusion 9
Glossary 10
References 10
Revision History 12
Licence and Disclaimer 12

© 2026 Darkwave Trading Limited, trading as Velocity Edge · CC BY 4.0

Executive Summary

Part 1 closed by asking why an agent given a rule stops following it hours later. The cited studies keep an instruction or a task fixed. They find that changing the conditions around it can move behaviour. They do not find this on every model or task. The evidence points less to how a rule was written than to the conditions it is read under. One study partly controls for rephrasing [1]. None, in our reading, compares wording with conditions more broadly.

Three conditions change over a long session. The context gets longer. The rule falls further behind the point where the model proposes. And part of the context may be replaced by a summary, which this series calls compaction, so that the rule is no longer there word for word. What the context contains changes too, and the hypothesis leaves that out.

The evidence is partial. For the second condition we know of no direct evidence at the distances a long session reaches. In our reading, outside recent work on compaction, much of the evidence measures accuracy on a task, not whether a rule was kept.

From this we state a hypothesis that we have not fitted. It covers a rule the model must keep by reading it, and whose breaches can be observed, within a model's stated context length. We hypothesise that such a rule decays with the session as each of those three conditions grows, other things equal. A rule presented first may gain from being first. The hypothesis makes no claim about distance for a rule presented only at the start, as a system prompt often is.

Some evidence runs against the hypothesis. Where a model judge found that the summary still carried a policy, one preprint found no violations, as in full context [2]. Those 90 cases were pooled with adversarial conditions. In our reading, the sample could not show a small rise. In another study, adherence fell when later turns pressed for a more novel, stronger or larger design, not over neutral turns [3]. Its neutral turn may have acted as weak pressure. Its judge, a model, missed most of the violations people found. In one rule-keeping test, on two models, the author attributes most of a fall to added content rather than length [4]. Its rules were formatting rules, and its padding was not exactly length-matched.

Work of many steps can be run as a chain of short contexts, each built fresh from a written record. A fresh context brings length and compactions back down, until it grows long in turn. One account shows the chain in use. One confounded measurement bears on its effect. We judge that this evidence does not yet show that lifespan matters apart from size.

1. The Question

Part 1, Context Is Configured, modelled an agent's context as the output of code. It called the action a model puts forward its proposal.

A rule as written does not change over a session. Where it sits, how much surrounds it, and whether it is still there word for word can all change.

© 2026 Darkwave Trading Limited, trading as Velocity Edge · CC BY 4.0

Product behaviours cited in this part are as read on 16 November 2026. Each may change with a release.

2. What the Evidence Shows

2.1 Length

Du and others report that performance fell as input grew [5]. They tested five models on tasks in mathematics, question answering and coding, with up to about 30,000 tokens added, within the models' stated context lengths. Their abstract puts the fall between 13.9 and 85 per cent. It says the fall persisted when the added tokens were only whitespace.

The abstract does not say which models and conditions that range covers. The stronger controls were run on the two smaller open models only. There the fall remained when the authors checked retrieval, hid the irrelevant tokens from attention, and placed the evidence just before the question. Retrieval was not perfect at the longest input. One closed model improved on one task at every length of added whitespace. Every experiment was run once.

2.2 Position

Liu and others found that several models of 2023 often did best when the needed information sat at the start or end of a long input. They did worse when it sat in the middle [6]. See also Part 1, Appendix A.5.

The pattern depended on the model and the task [6].

Later work suggests the effect may be weaker or absent in some newer models, on the tasks tested [7], [8].

That later work has limits. One study tested one newer model, on 26 simple factual questions it answered correctly at every size it could test. In our reading, a small effect could not show there. The other used mainly a classification task. By its authors' reading of one figure, from one run, one of four models favoured the start.

2.3 Turns

In one large simulation by Laban and others, giving an instruction in pieces over turns cost reliability. Across six generation tasks and 15 models, performance fell by 39 per cent on average against giving the instruction whole [1]. By the authors' measures, the loss was in large part a rise in unreliability. Aptitude fell by 16 per cent on average. They call that fall non-significant.

The conversations were simulated. The authors say they are not representative of natural conversation. They think the fall they measured is most likely an underestimate.

In our reading, the loss came with an instruction that started incomplete. Giving all the pieces in one turn, still rephrased, kept 95.1 per cent of the whole-instruction score on average. That condition partly controls for rephrasing. When the pieces arrive is a matter of timing, one of Part 1's settings. The design does not separate starting incomplete from being spread over turns.

© 2026 Darkwave Trading Limited, trading as Velocity Edge · CC BY 4.0

Li and others found, in 2024, significant drift from a system prompt within eight rounds of dialogue between two instructed chatbots [9]. That was on one older model. In an appendix, a second older model kept to its prompt better but still drifted. The authors hypothesise that decaying attention to the prompt may contribute to drift.

2.4 What These Studies Do Not Measure

Of the studies above, in our reading, only Li's measures whether an instruction was still followed, on two older models in synthetic dialogues.

3. Failures and the Settings They Belong To

Table 1 assigns reported failures to a setting of Part 1. The first four rows rest mainly on studies of models, not of long-running agents. That these failures grow over an agent's session is our inference.

Table 1: Reported failures and the setting we assign each to. Sources: as cited in each row.

Failure	What is reported	Setting
Length	Performance can fall as input grows [5], [10]. An Anthropic post of September 2025 describes an attention budget that every added token draws on, as an analogy [11] (Part 1, Appendix A.4)	Size, composition
Positional loss	Several models of 2023 often did worse when the needed material sat in the middle of a long input. The pattern depended on the model and the task [6]	Position
Drift over turns	One older model drifted significantly from a system prompt within eight rounds; a second drifted less [9]	Timing, lifespan
Instruction given in pieces	In simulation, reliability fell when an instruction was revealed in pieces over turns. In our reading, the loss came with an instruction that started incomplete. The design does not separate that from being spread over turns [1]	Timing
Loss in compaction	That Anthropic post warns that overly aggressive compaction can lose context whose importance shows only later [11]. A preliminary study on one benchmark reports that compression, the term it mainly uses, can weaken the influence of recent interactions [12]	Lifespan
History eviction	In one descriptive test on one model, success was far lower when old messages were evicted, and keeping the plan with a few recent messages did not recover the gap. Its authors say the result cannot be attributed to removing the plan [13]	Lifespan, timing

© 2026 Darkwave Trading Limited, trading as Velocity Edge · CC BY 4.0

Failure	What is reported	Setting
Anchoring	In optimisation tasks on one model, agents that accumulated history became anchored to their own past attempts, in the authors' reading [14]	Lifespan
Hidden stop signals	In work for a workshop of December 2025, summarising lengthened runs for two models, compared with masking. The authors suggest summaries may hide signs that the agent should already stop [15]	Lifespan

4. A Hypothesis of Decay

Take a rule *r* that the model must keep by reading it, and whose breaches can be observed. Let *q_r*(*t*) be the probability that the model's proposal at a step *t* where *r* applies breaks it. We hypothesise:

$$\text{logit } q_r(t) = \alpha_r + \beta \cdot L(t) + \gamma \cdot \delta_r(t) + \eta \cdot k_r(t)$$

L(*t*) is the length of the context at step *t*, in tokens. Part 1, Appendix A.4, conjectures how length could weaken rule-keeping. Its mechanisms are plausible, not shown. On two open models, Du's fall remained with the added tokens hidden from attention. That suggests attention dilution alone does not explain it. *δ_r*(*t*) is the distance in tokens from the rule's most recent word-for-word presentation still in the context to the proposal (Part 1, Appendix A.5). *k_r*(*t*) is the number of compactions since the rule was last presented word for word (Part 1, Appendix A.6). A compaction that keeps the rule word for word is treated as presenting it again. If no presentation remains, *δ_r* runs from where the last one stood, in the tokens now in the context. *k_r* then includes the compaction that removed it. A rule lost without compaction is outside the hypothesis. *α_r* is the rule's base difficulty in a short context. *β*, *γ* and *η* are weights for a given model. We hypothesise that each is positive, and we fit none. We call this the decay hypothesis.

Length has the most evidence of the three, mostly of task accuracy (Section 2.1) [5], [10]. In the one rule-keeping test we cite that set length against content, on two of its twelve models, the author attributes most of the fall to content [4].

For distance we know of no direct evidence on rule-keeping at the distances a long session reaches, under a model's normal attention. In one study, attention to a goal's tokens fell over 50 turns. Hiding all but the latest 4,096 tokens from attention cut recall in the models tested, sharply in most [16]. But turns and distance rose together. The authors treat the intervention as out of distribution for two of the four models. One of those models' compliance with a policy, in a separate task, barely changed under that intervention. By analogy, Liu's result for evidence at the start of an input suggests that the effect of distance need not rise steadily [6].

Two cited preprints of 2026 bear on compaction, mostly on what follows a single compaction.

© 2026 Darkwave Trading Limited, trading as Velocity Edge · CC BY 4.0

In one, prohibited tool actions rose from none in full context to 30 per cent of episodes after compaction under a tight word budget, on its author's benchmark across seven models [2]. A model judge sorted the episodes, on five models and pooled with adversarial conditions. Violations were 38 per cent where the summary had dropped the policy. They were none where the judge found the summary still carried it. The preprint also says that violations compounded over repeated compactions, from none to 78 per cent over four rounds. Its main text gives no models or episodes for this, so for the effect of repeated compactions the evidence is thin.

The other mainly measured whether a rule survived [17]. By its abstract, the methods tested kept 17 per cent of injected user constraints on average. In its motivating test, on one model, 28 of the 30 sessions with unapproved deletions had lost the constraint in compaction.

Chen's "none" covers summaries that still carried the policy, perhaps in paraphrase. k_r still covers such a rule, and for it this evidence suggests little or no rise. A keyword check on seven models, pooled with adversarial conditions, gave 1 per cent where it survived, 43 where dropped. In our reading, the judge's 90 cases were too few to show a small rise. Both preprints bear on losing the rule more than on the number of compactions.

The hypothesis has four limits, and a scope. First, it is claimed only within a model's stated context length. Second, its linear form is a simplification. A rule presented first may gain from being first. For a rule placed first and never presented again, distance and length move together, so the hypothesis makes no separate claim about distance for it. Third, it gives each rule its own difficulty. But it gives every rule of a model the same rate of fading on the log-odds scale. It has no term for whether fading levels off. Fourth, it has no term for what the context comes to contain, or for how many rules it carries. Like Part 1's model, it is of transformer-based language models (Part 1, Section 2).

In our reading, some evidence runs against length and distance alone.

In one preprint on research ideation, adherence to constraints did not fall over six neutral turns [3]. Adherence fell, in most of seven models, when later turns pressed for a more novel, stronger or larger design. Six of the seven could still restate the constraints fully. The author reads this as a challenge to forgetting as the cause.

Its author calls it descriptive and its neutral turn not truly neutral. Its judge, a model, missed most of the violations people found.

In our reading, other work points the same way [4], [18].

Of the settings, size moves L. Timing and the rule's position move δ_r. Lifespan moves k_r. A fresh context starts L at the length of the rules and the record. L then grows within it.

Repeating a rule just before the proposal brings δ_r near zero. Presenting the rules again word for word after a compaction brings k_r to zero. The hypothesis predicts a lower violation rate after either when the fall in the γ or η term exceeds β times the tokens added to L.

© 2026 Darkwave Trading Limited, trading as Velocity Edge · CC BY 4.0

Repeating the system prompt did well late in a dialogue, on one model [9]. It used up much of the context.

Any test keeps the wording fixed. What makes up the added length is fixed too, or varied as its own factor: neutral padding against task content. A margin is set before the test. Violations are judged by people or by code, and steps are grouped by run in the analysis. We are wrong about a term, for a model, if the log-odds of a rule's violation do not rise with that term by more than the margin. That is judged at steps where the rule applies, with enough runs to show it, and with the other two terms fixed or modelled. If it is so of any one term, the hypothesis as stated is wrong for that model.

The test that would fit the hypothesis varies three things: the length of the context, the distance of the rule's last presentation, and the number of compactions. Length and distance separate only when the rule's presentation moves independently of length, as when it is presented again partway through.

5. A Chain of Short Contexts

Work of many steps can run as a chain of short contexts, each built fresh from a written record. Each context in the chain is a link.

Slack's engineering post of April 2026 says its security investigation service passes no message history from one agent invocation to the next [19]. It reports no measure of the design's effect.

In one preprint, Tan and others ran a team of four agents that periodically discarded its search state and restarted from the best candidate so far [14]. At the same budget, it scored up to 15 per cent above four parallel agents and up to 39 per cent above a single agent.

The team shared candidates between its agents, which the parallel agents did not, so the gain over them mixes sharing with forgetting. The preprint used one model, three runs, four tasks and no significance test. The 39 per cent, over a single agent, was on one task. On another, the gain over parallel agents was about 2 per cent.

One test bears on what a record must keep. In it, success was far lower when old messages were evicted [13]. Its authors say this cannot be attributed to removing the plan. Keeping the plan did not close the gap. That test used no chain built from a written record.

We judge that the evidence for the chain does not yet show that lifespan matters apart from size.

Every link presents the rules word for word and takes the work's state from the record. What the record leaves out of earlier work is lost to that link. In that sense a record is a running summary. We know of no study that compares a chain with compaction followed by presenting the rules again word for word.

© 2026 Darkwave Trading Limited, trading as Velocity Edge · CC BY 4.0

6. Compaction as a Last Resort

Where a context must stay alive, we propose an order of steps. No study we know of has tested it whole. Summarising comes last (Part 1, Appendix A.6):

1. **Clear** tool results already used and recorded.
2. **Mask** observations older than a fixed number of recent turns, each replaced by a placeholder that keeps its status, identifiers and pointer.
3. **Offload** large tool results to files, behind previews that keep the label of where each came from.
4. **Point**: replace old tool inputs already saved with pointers to them.
5. **Summarise**, only when the steps above do not free enough room.

One research blog post, on work for a workshop of December 2025, compared masking with summarising on cost and task success [15]. Masking often matched or slightly beat summarising at solving the tasks.

That comparison measured no rule-keeping, and used a plain placeholder, not step 2's.

We put summarising last because we judge that what it removes is hardest to restore. One preliminary study reports that compression can bring more blocked actions, repeated exploration and instability across runs [12]. The same study reports plain truncation matching or slightly beating summarising at 16K, 8K and 4K tokens. At the budget it calls the tightest, it reports summarising far ahead in mean pass rate: 72.8 against 42.2 per cent. Its text does not name that budget. In our reading, step 5 would apply there. There summarising is compared with truncation, not steps 1 to 4. The study's design let the agent query the state again.

7. What a Chain Does Not Fix

We argue that where two chains run side by side, no single context sees a relation between them unless it reads both records. A misreading that every fresh context shares recurs in every link.

Conclusion

A rule that the model keeps only by reading it is carried by the context it was read in. Over a long session that context changes. The evidence that the change matters is partial. It is broadest for length. It is mixed for the position of evidence. For distance, we know of no direct test at the distances a long session reaches. For compaction the evidence is recent, and two preprints we cite bear on losing rules, mainly after one compaction. Elsewhere, in our reading, much of it measures task accuracy, not rule-keeping.

In our reading of two cited studies of rule-keeping, what the context came to contain also appeared to matter [3], [4]. Some evidence runs against the hypothesis (Section 4). The evidence does not say why a rule stops being followed.

A fresh context brings length and compactions back down for a time. A chain of short contexts supplies one at every link. One cost is what each record leaves out.

© 2026 Darkwave Trading Limited, trading as Velocity Edge · CC BY 4.0

If a rule kept by reading can fade, which rules need not be kept that way at all?

Glossary

Term	Meaning
Chain of short contexts	Work of many steps run as a sequence of contexts, each built fresh from a written record, in place of one context kept alive
Compaction	Replacing part of a context with a summary so that the work can go on in the same context; what was summarised is no longer present word for word
Decay hypothesis	This part's unfitted hypothesis that, within a model's stated context length, a rule kept by reading, whose breaches can be observed, becomes more likely to be broken as the context lengthens, as the rule's last presentation falls further behind the proposal, and with each compaction since it was last present word for word; a rule presented first may gain from being first, and no distance claim is made for a rule presented first and never again
Distance	The number of tokens from a rule's most recent word-for-word presentation still in the context, or, if none remains, from where the last one stood, to the model's proposal
Drift	An agent's following of its rules weakening over a session
Link	One context in a chain of short contexts
Masking	Here, observation masking, not Part 1's masking of tool choice: replacing an older observation with a placeholder. In Section 6's order of steps the placeholder keeps the observation's status, identifiers and pointer. The comparison cited there used a plain placeholder
Observation	What a tool returns into the context
Rule kept by reading	A rule whose only support is that the model reads it in its context and acts on it
Truncation	Dropping the oldest part of a context without summarising it

References

[1] P. Laban, H. Hayashi, Y. Zhou, and J. Neville, "LLMs get lost in multi-turn conversation." arXiv:2505.06120, 2025. Accessed: Nov. 15, 2026. [Online]. Available: https://arxiv.org/abs/2505.06120

[2] S. Chen, "Governance decay: How context compaction silently erases safety constraints in long-horizon LLM agents." arXiv:2606.22528, version 2, 27 June 2026, 2026. Available: https://arxiv.org/abs/2606.22528

[3] G. Kruthof, "Models recall what they violate: Constraint adherence in multi-turn LLM ideation." arXiv:2604.28031, version 2, 4 May 2026, 2026. Available: https://arxiv.org/abs/2604.28031

© 2026 Darkwave Trading Limited, trading as Velocity Edge · CC BY 4.0

[4] Y. Gamage, "Omission constraints decay while commission constraints persist in long-context LLM agents." arXiv:2604.20911, version 1, 22 April 2026, 2026. Available: https://arxiv.org/abs/2604.20911

[5] Y. Du *et al.*, "Context length alone hurts LLM performance despite perfect retrieval." arXiv:2510.05381; Findings of EMNLP 2025, 2025. Accessed: Nov. 15, 2026. [Online]. Available: https://arxiv.org/abs/2510.05381

[6] N. F. Liu *et al.*, "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024, doi: 10.1162/tacl_a_00638.

[7] M. McKinnon, "Retrieval quality at context limit." arXiv:2511.05850, 2025. Accessed: Nov. 15, 2026. [Online]. Available: https://arxiv.org/abs/2511.05850

[8] N. Esmi, M. Nezhad-Moghaddam, F. Borhani, A. Shahbahrami, A. Daemdoost, and G. Gaydadjiev, "GPT-5 vs other LLMs in long short-context performance." arXiv:2602.14188, 2026. Accessed: Nov. 15, 2026. [Online]. Available: https://arxiv.org/abs/2602.14188

[9] K. Li *et al.*, "Measuring and controlling instruction (In)stability in language model dialogs." arXiv:2402.10962; COLM 2024, 2024. Accessed: Nov. 15, 2026. [Online]. Available: https://arxiv.org/abs/2402.10962

[10] K. Hong, A. Troynikov, and J. Huber, "Context rot: How increasing input tokens impacts LLM performance." Chroma Technical Report, 14 July, 2025. Accessed: Nov. 15, 2026. [Online]. Available: https://www.trychroma.com/research/context-rot

[11] P. Rajasekaran, E. Dixon, C. Ryan, and J. Hadfield, "Effective context engineering for AI agents." Anthropic, 29 September, 2025. Accessed: Nov. 15, 2026. [Online]. Available: https://www.anthropic.com/research/effective-context-engineering-for-ai-agents

[12] G. Min, L. Wu, M. Darbari, C. Chen, and L. Hong, "Towards reliable context compression for long-horizon agents: An empirical study of execution instability." arXiv:2608.06503 (v2, 14 October 2026), 2026. Accessed: Nov. 15, 2026. [Online]. Available: https://arxiv.org/abs/2608.06503

[13] A. Mehta and A. Datta, "Plans fade fast: Why context management is load-bearing for LLM agents." arXiv:2606.22953 (v2, 12 October 2026), 2026. Accessed: Nov. 15, 2026. [Online]. Available: https://arxiv.org/abs/2606.22953

[14] S. Tan *et al.*, "Scaling discovery requires forgetting." arXiv:2610.20563, 2026. Accessed: Nov. 15, 2026. [Online]. Available: https://arxiv.org/abs/2610.20563

[15] T. Lindenbauer *et al.*, "Cutting through the noise: Smarter context management for LLM-powered agents." JetBrains Research blog, December; NeurIPS 2025 Deep Learning 4 Code workshop, 2025. Accessed: Nov. 15, 2026. [Online]. Available: https://blog.jetbrains.com/research/2025/12/efficient-context-management/

[16] V. Dongre, J. Hsieh, V. D. Lai, S. Yoon, T. Bui, and D. Hakkani-Tür, "When attention closes: How LLMs lose the thread in multi-turn interaction." arXiv:2605.12922, version 2, 6 October 2026, 2026. Available: https://arxiv.org/abs/2605.12922

© 2026 Darkwave Trading Limited, trading as Velocity Edge · CC BY 4.0

[17] Z. Wang, Y. Zhang, D. Lee, N. Baracaldo, and Y. Yang, “Lost in context compaction: When LLM systems drop side constraints.” arXiv:2608.11242, version 2, 11 October 2026, 2026. Available: https://arxiv.org/abs/2608.11242

[18] A. Sinha, A. Arun, S. Goel, S. Staab, and J. Geiping, “The illusion of diminishing returns: Measuring long horizon execution in LLMs.” arXiv:2509.09677, version 3, 13 March 2026, 2025. Available: https://arxiv.org/abs/2509.09677

[19] D. Marks, “Managing context in long-run agentic applications.” Slack Engineering, 13 April, 2026. Accessed: Nov. 15, 2026. [Online]. Available: https://slack.engineering/managing-context-in-long-run-agentic-applications/

Revision History

Version	Date	Change
1.0.0	2026-11-19	First publication
1.0.1	2026-11-20	Part 1 named by its new title, Context Is Configured

Licence and Disclaimer

© 2026 Darkwave Trading Limited, trading as Velocity Edge. This paper is licensed under the Creative Commons Attribution 4.0 International licence (CC BY 4.0, https://creativecommons.org/licenses/by/4.0/): you may share and adapt it for any purpose, provided you credit Phillip Pienaar and Velocity Edge, keep this copyright notice, link to the licence, link to the original where reasonably practicable, and indicate any changes. The Darkwave and Velocity Edge names and logos are not covered by this licence.

This paper is for education and information. It is not investment advice, and nothing in it is a recommendation to trade.