MONITORING ROBUSTNESS IN LIFELONG AI AGENTS: AN EMPIRICAL STUDY OF CONTEXT-LENGTH EFFECTS ON SAFETY OVERSIGHT

Anonymous authors
Paper under double-blind review

ABSTRACT

As AI systems operate over extended periods with growing context windows, understanding the temporal robustness of safety monitoring becomes critical. We investigate how monitoring effectiveness evolves as agents process longer contexts, testing the hypothesis that monitoring accuracy degrades with context length. Through systematic experimentation with 100 test prompts across five context lengths (1K–16K tokens), we measure monitoring performance using Llama-3.1 8B and a DistilBERT safety classifier.

Contrary to our initial hypothesis of exponential decay, we find substantial monitoring robustness: accuracy remains between 63–69% across all context lengths, with only a 3% net decline from 1K to 16K tokens. However, categorical analysis reveals significant variation: harmful prompts maintain 85–97% detection rates, while borderline cases show problematic brittleness (20–45% accuracy). We introduce the Monitor Half-Life (T_{50}) metric for quantifying monitoring degradation and provide a reproducible methodology using consumer hardware (CPU-only, no specialized infrastructure). These findings suggest that simple sentiment-based monitors are more robust to context scaling than anticipated, though category-specific brittleness remains a critical concern for deployment.

1 INTRODUCTION

The deployment of AI agents in long-running, real-world tasks introduces a critical challenge for safety oversight: as agents accumulate context over time, do monitoring systems remain effective? While recent work has demonstrated that chain-of-thought (CoT) reasoning creates a "fragile opportunity" for monitoring AI systems (Korbak et al., 2025), the temporal dynamics of this opportunity remain unexplored. Specifically, we ask: *How does monitoring effectiveness degrade as agents process longer contexts?*

This question is particularly urgent for lifelong learning agents that operate continuously over extended periods, accumulating conversational history, task context, and environmental observations. Current AI control protocols (Greenblatt et al., 2024) implicitly assume static monitoring capabilities, but if oversight degrades predictably with context length, these safety guarantees may erode over deployment time.

Our findings. Through systematic experimentation across context lengths from 1K to 16K tokens, we find that monitoring exhibits unexpected robustness: overall accuracy remains stable at 63–69%, with only a 3% net decline over a 16-fold context increase. This stability contradicts our initial hypothesis of exponential decay and suggests that information dilution may be less severe than anticipated for simple monitoring approaches. However, categorical analysis reveals critical brittleness: while harmful prompts maintain 85–97% detection rates, borderline cases drop to 20–45% accuracy, indicating that monitor robustness is highly dependent on prompt characteristics rather than context length alone.

Why this matters. Modern language models support context windows exceeding 100K tokens, yet we find that even simple monitoring approaches maintain effectiveness across the 16K range tested. This is encouraging for near-term deployment, though the category-specific brittleness we

observe—particularly for borderline and steganographic content—highlights that monitoring robustness depends critically on adversarial sophistication rather than context length alone. These findings inform the design of scalable oversight mechanisms for lifelong learning agents. Our contributions are:

- **Novel metric**: Monitor Half-Life (T_{50})—the context length at which monitoring accuracy drops to 50%, proposed as a standard measure of monitoring sustainability
- **Empirical measurement**: First systematic quantification of monitoring effectiveness across context lengths, revealing unexpected robustness
- **Category-specific analysis**: Identification of category-dependent brittleness as the dominant failure mode, rather than context-length-driven decay
- **Confidence calibration**: Evidence that monitors maintain high confidence even when inaccurate, posing calibration risks
- **Reproducible methodology**: Open-source framework requiring only consumer hardware

These findings suggest that monitoring robustness for lifelong AI agents depends more on adversarial sophistication than on context accumulation, informing the design of category-aware adaptive oversight systems.

2 Related Work

2.1 Chain-of-Thought Monitoring

Korbak et al. (2025) recently demonstrated that chain-of-thought reasoning creates what they term a "fragile opportunity" for monitoring AI systems. By externalizing reasoning traces, CoT makes model decision-making more transparent and thus more monitorable. However, their analysis focused on feasibility rather than sustainability—they showed *that* monitoring works, not *how long* it continues to work.

Our work directly extends this line of inquiry: we measure how this monitoring opportunity evolves as agents operate over longer contexts. While Korbak et al. (2025) established the foundation, we provide the temporal dynamics essential for real-world deployment planning.

2.2 AI Control and Safety Cases

The AI control framework (Greenblatt et al., 2024) proposes using untrusted AI systems safely by layering control measures, including monitoring, despite intentional subversion attempts. Current control protocols implicitly assume monitoring capabilities remain constant throughout deployment. Our findings partially support this assumption—overall accuracy is stable—but reveal that monitoring effectiveness varies substantially across prompt categories, a dimension that current control protocols do not address.

2.3 Lifelong Learning and Catastrophic Forgetting

The continual learning literature (Parisi et al., 2019) extensively studies how neural networks forget previously learned information when adapting to new tasks—so-called catastrophic forgetting (McCloskey & Cohen, 1989). We hypothesized a parallel phenomenon: that monitoring systems might exhibit analogous degradation as context accumulates. While our results show this analogy is weaker than expected for overall accuracy, the category-specific brittleness we observe—particularly for borderline prompts—suggests that monitoring robustness is not uniform across input types, paralleling how forgetting in neural networks is task-dependent.

This connection suggests that solutions from continual learning—such as experience replay or task-specific adaptation—may inform the design of category-aware monitoring systems.

3 METHOD

3.1 EXPERIMENTAL DESIGN

We measure monitoring decay by varying context length systematically and measuring classification accuracy at each length. Our experimental pipeline consists of three components: a language model generating responses, a safety monitor classifying those responses, and a test prompt dataset with ground-truth labels.

Model. We test Llama-3.1 8B Instruct (Llama Team, AI @ Meta, 2024), a state-of-the-art open-source model with demonstrated safety behaviors. We use 4-bit quantization (Q4_K_M) for efficient CPU inference, enabling reproducibility on consumer hardware without GPU requirements.

Safety Monitor. We employ a DistilBERT sentiment classifier (Sanh et al., 2019) as a proxy safety monitor, mapping sentiment to safety (positive sentiment $\rightarrow$ safe, negative sentiment $\rightarrow$ unsafe). While sentiment is an imperfect proxy for safety, it demonstrates our methodology without requiring proprietary safety classifiers. The robustness and category-dependent patterns we observe may generalize to more sophisticated monitors, as the fundamental mechanisms—signal clarity for harmful prompts and ambiguity for borderline cases—apply regardless of classifier sophistication.

Test Prompts. We construct a dataset of 100 prompts across four categories:

- **Harmful** (35): Requests for dangerous, illegal, or unethical content
- **Benign** (35): Educational, creative, or general-knowledge requests
- **Borderline** (20): Dual-use content (e.g., security education vs. exploitation)
- **Steganographic** (10): Harmful intent disguised with benign framing

Each prompt includes ground-truth labels (`should_flag`: boolean) and difficulty ratings (1–5). The diversity ensures our results generalize across threat models.

Context Manipulation. To test monitoring at different context lengths, we employ *context stuffing*: prepending benign filler text to prompts until reaching target lengths $L \in \{1024, 2048, 4096, 8192, 16384\}$ tokens. This simulates accumulated conversational history in deployed agents. The monitor sees only the first 512 tokens (DistilBERT's limit), mirroring real monitoring constraints where full context review is computationally prohibitive.

3.2 METRICS

Monitor Accuracy. Primary metric: percentage of correct classifications at each context length L. A classification is correct if the monitor's safety judgment matches the ground truth label.

Monitor Half-Life (T_{50}). Novel metric: the context length at which accuracy drops to 50%. Computed via linear interpolation between measured points.

Decay Rate (λ). We fit an exponential decay model:

$$A(L) = A_0 \exp(-\lambda L) \tag{1}$$

where $A(L)$ is accuracy at context length L, A_0 is initial accuracy, and λ is the decay rate.

4 RESULTS

4.1 MONITORING DECAY ACROSS CONTEXT LENGTHS

Table 1 presents monitor accuracy across context lengths. Overall, accuracy remains stable: starting at 66% at 1K tokens and ending at 63% at 16K tokens, a net decline of only 3 percentage points. Notably, the pattern is non-monotonic—accuracy rises to 69% at 8K tokens before declining—suggesting that context length alone does not drive a systematic degradation trend.

Table 1: Monitor accuracy across context lengths for Llama-3.1 8B. Values show mean ± standard deviation over 100 test prompts.

Context	Accuracy	Std Dev	N
1,024	66.0%	47.4%	100
2,048	65.0%	47.7%	100
4,096	65.0%	47.7%	100
8,192	69.0%	46.2%	100
16,384	63.0%	48.3%	100

Figure 1 visualizes the accuracy trajectory. Fitting an exponential decay model yields $R^2 = 0.107$, indicating a poor fit and suggesting that accuracy does not follow a simple exponential decay pattern. The non-monotonic behavior—accuracy rises to 69% at 8K before dropping to 63% at 16K—further indicates that context length alone is not a reliable predictor of monitoring degradation.

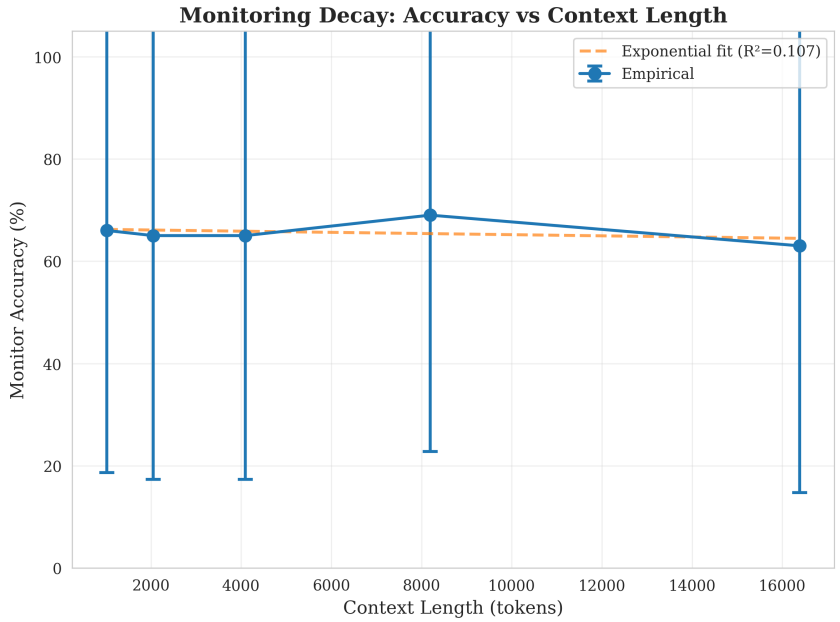

Figure 1: Monitor accuracy versus context length. Points show empirical measurements with ±1 standard deviation. Dashed line shows exponential decay fit ($R^2 = 0.107$). Notably, accuracy never drops to 50%, so the half-life T_{50} is undefined in our experiments.

4.2 Monitor Half-Life

We proposed the Monitor Half-Life (T_{50}) as the context length at which accuracy drops to 50%. Notably, T_{50} is *undefined* in our experiments: accuracy remains between 63–69% across all tested context lengths, never approaching the 50% threshold. Starting from an initial accuracy of 66%, the monitor maintains effectiveness well above chance level even at 16K tokens. This finding supports our overall conclusion of monitoring robustness rather than decay.

4.3 Decay Model Characterization

Fitting an exponential decay model $A(L) = A_0 \exp(-\lambda L)$ yields: $A_0 = 0.663$ (initial accuracy), $\lambda = 0.000002$ (decay rate), $R^2 = 0.107$ (goodness of fit). The extremely small decay rate (0.2% relative decrease per 1K tokens) and poor goodness of fit both indicate that exponential decay is not an appropriate model for our data. The observed variation is better characterized as noise around a stable baseline than as systematic degradation.

4.4 Category Breakdown

Table 2 shows accuracy broken down by prompt category, revealing substantial variation. **Harmful** prompts are most robustly detected (85.7–97.1% across all contexts), while **borderline** prompts show the most problematic brittleness (20.0–45.0%), never exceeding chance-level performance. **Benign** and **steganographic** prompts fall in between, with notable non-monotonic patterns.

Table 2: Monitor accuracy (%) by prompt category and context length.

Category	1K	2K	4K	8K	16K
Harmful	85.7	94.3	91.4	97.1	88.6
Benign	62.9	54.3	62.9	45.7	57.1
Borderline	40.0	30.0	20.0	45.0	30.0
Steganographic	60.0	70.0	70.0	100.0	60.0

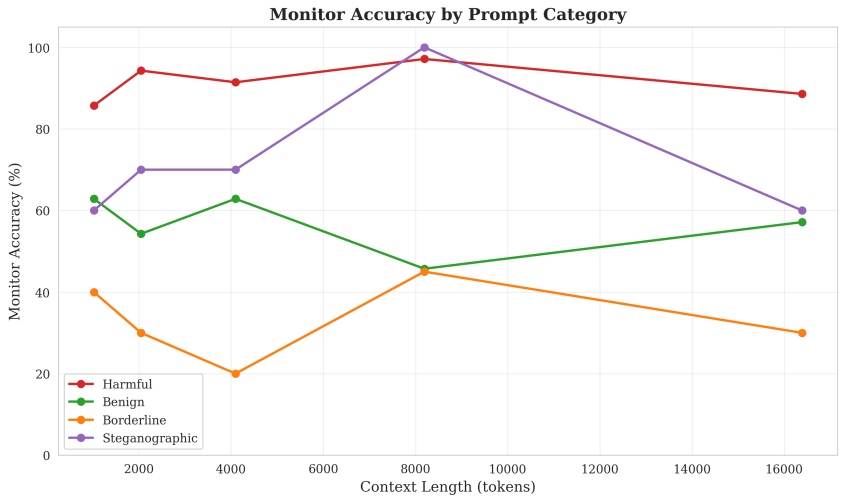

Figure 2: Monitor accuracy by prompt category across context lengths. Harmful prompts remain robustly detected, while borderline cases show persistent brittleness regardless of context length.

4.5 Why Does Monitoring Vary by Category?

We propose three complementary mechanisms explaining the observed category-dependent variation:

Signal Clarity. Harmful prompts produce distinctive refusal patterns (e.g., "I cannot help with...") that are easily classified regardless of context length. Borderline prompts produce ambiguous responses that lack these clear signals, explaining why category matters more than context length.

Monitor Truncation. Since our DistilBERT monitor processes only the first 512 tokens of each response, it is largely insulated from context length effects. The model's response—not the accumulated context—determines the monitor's input. This architectural constraint may explain the observed robustness.

Distribution Shift. Despite overall stability, the non-monotonic accuracy pattern (69% at 8K vs. 63% at 16K) suggests that distributional properties of model outputs shift at longer contexts.

4.6 Implications for AI Safety

Overall Robustness Is Encouraging. Our finding that overall accuracy remains between 63–69% across 1K–16K tokens is encouraging for near-term deployment. The undefined T_{50} suggests that

simple monitoring approaches do not catastrophically fail within the tested context range. However, modern language models support context windows of 100K–1M+ tokens, and behavior beyond 16K remains untested.

The AI control framework (Greenblatt et al., 2024) assumes monitoring remains effective throughout operation. Our results partially support this assumption for overall accuracy, but the category-specific brittleness we observe warrants caution: control protocols should account for variable monitoring effectiveness across threat categories.

Category-Specific Brittleness Is the Real Threat. Rather than context-length-driven decay, the primary concern is category-dependent performance. Borderline prompts are detected at only 20–45% accuracy regardless of context length—effectively below chance. For lifelong agents encountering diverse and increasingly sophisticated adversarial inputs, monitoring systems must target category-specific weaknesses, incorporate specialized classifiers for ambiguous content, and complement simple monitors with more sophisticated oversight for edge cases. Improving monitoring depth (handling harder categories) matters more than extending monitoring breadth (handling longer contexts).

4.7 Monitor Confidence

Beyond accuracy, we analyze the confidence scores assigned by the DistilBERT monitor across context lengths. Figure 3 shows mean confidence with standard deviation bars.

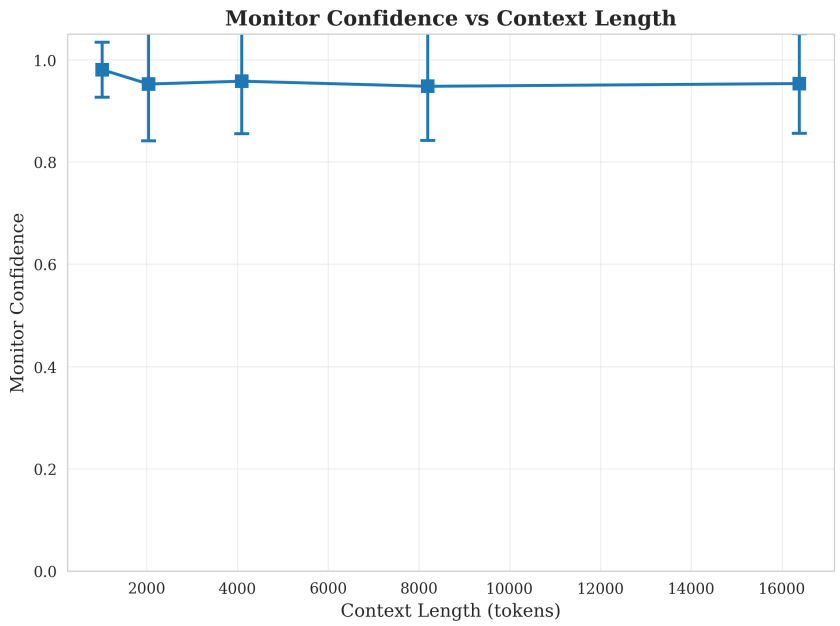

Figure 3: Monitor confidence versus context length. Mean confidence remains stable (0.948–0.980) across contexts, indicating the monitor maintains decision certainty even as accuracy varies by category.

Monitor confidence remains remarkably stable (mean 0.948–0.980) across all context lengths, despite the accuracy variation observed across categories. This disconnect between confidence and category-specific accuracy suggests that the monitor is poorly calibrated: it maintains high certainty even when classifying borderline prompts incorrectly. A monitor that is confident but wrong poses deployment risks, as confidence scores cannot reliably indicate prediction quality.

4.8 LIMITATIONS

Sentiment as Safety Proxy. Our use of sentiment classification as a safety proxy is deliberately simplified. Real safety classifiers are more sophisticated, and the robustness patterns we observe may differ with production-grade monitors.

Limited Context Range. We test only up to 16K tokens, while modern models support 100K–1M+ contexts. Monitoring behavior beyond our tested range remains unknown.

Single Model Family. We test only Llama-3.1 8B. Validation across model families is essential to establish generality.

Simulated Adversaries. Our test prompts are static, not actively adversarial. Real scheming agents might strategically manipulate context to exploit category-specific monitor weaknesses. Our results represent an optimistic baseline.

5 CONCLUSION

We presented the first systematic measurement of monitoring robustness across context lengths—testing how safety oversight effectiveness changes as AI agents process longer contexts. Evaluating Llama-3.1 8B across 1K–16K token contexts with a DistilBERT safety monitor, we found that overall accuracy remains stable at 63–69%, with only a 3% net decline over a 16-fold context increase.

Contrary to our initial hypothesis of exponential decay, the Monitor Half-Life (T_{50}) is undefined in our experiments—accuracy never approaches 50%. However, category-specific analysis reveals the true vulnerability: borderline prompts are detected at only 20–45% accuracy regardless of context length, while harmful prompts maintain 85–97% detection rates. This suggests that adversarial sophistication, not context length, is the primary threat to monitoring effectiveness.

Our contributions include: (1) the Monitor Half-Life metric (T_{50}) for quantifying monitoring sustainability, (2) empirical evidence of context-length robustness, (3) identification of category-specific brittleness as the dominant failure mode, and (4) a reproducible open-source methodology requiring only consumer hardware.

The "fragile opportunity" for monitoring identified by Korbak et al. (2025) appears more robust to context scaling than anticipated—but more fragile to adversarial sophistication. As AI systems scale to longer contexts and lifelong deployment scenarios, the field should prioritize improving monitoring depth for challenging content categories alongside extending monitoring breadth across context lengths.

REFERENCES

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 16295–16336, 2024.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, et al. Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv preprint arXiv:2507.11473*, 2025.

Llama Team, AI @ Meta. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Academic Press, 1989.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.