THE IMPACT OF CONTEXT WINDOW CONSTRAINTS ON REACT AGENTS IN CRYPTOGRAPHIC CTF CHALLENGES

The 'More is Better' Fallacy.

HOW DOES SYSTEMATICALLY VARYING CONTEXT-WINDOW CONSTRAINTS INFLUENCE THE REASONING PERFORMANCE AND FAILURE MODES OF REACT STYLE AGENTS WHEN SOLVING CRYPTOGRAPHIC CTF CHALLENGES?

TU Delft

Author: Yusuf Kose
Contact: ybkose@tudelft.nl
Supervisor: Dr. Zeki Erkin
Examiner: Mitchell Olsthoorn

The Problem & The Gap

- **The Industry Assumption:** Current trends prioritize "Infinite Context" (1M+ tokens) under the assumption that maximizing history maximizes reasoning quality.
- **The Reality (The "Triple Penalty"):** Recent studies [1, 2] suggest that excessive context is a liability, not a feature depending on the environment. It introduces:
  - **Performance Degradation:** Loss of latent associations (Lost-in-the-Middle).
  - **Latency Bottlenecks:** Increasing Time-To-First-Token delays.
  - **Unjustified Cost:** Exponentially higher inference costs for diminishing returns.
- **The Gap:** Security benchmarks focus on capability (Can it solve X?) but ignore stability (Can it solve X efficiently?). There is no defined safe operating floor established for crypto-agents.

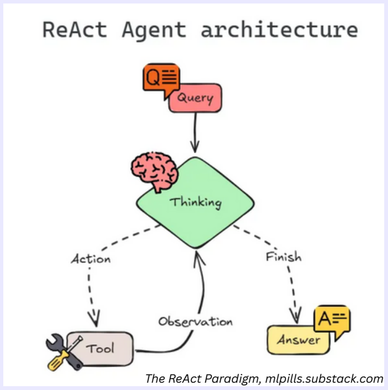
How do systematically varying context constraints influence the reasoning performance, efficiency, and failure modes of ReAct agents?

Methodology & Experimental Design

- **The Independent Variable:** Strictly enforced context budgets at 8k, 16k, 32k, and 64k tokens.

```
Experimental Budget Levels:
    - Severely constrained: 8,000 tokens
    - Moderately constrained: 16,000 tokens
    - Lightly constrained: 32,000 tokens
    - Control (unconstrained): 128,000 tokens
```

- **Agent Architecture:** ReAct Paradigm (Reason+Act), enforcing a strict *Thought* → *Action* → *Observation* execution loop to generate verbose reasoning traces.



The ReAct Paradigm, mlpills.substack.com

- **Truncation Policy:** Utilized deterministic FIFO (First-In-First-Out) truncation rather than semantic compression to isolate context capacity as the sole variable.
- **The Subject Models:** Evaluated DeepSeek-V3.2 (Open Weights), Gemini-3-Pro-Preview, and GPT-5.1 on 50 stratified cryptographic challenges from the AICrypto Dataset.
- **Dependent Variables (Metrics):**
  1. Success Rate (Binary Flag Capture)
  2. Failure Taxonomy (Hallucination vs. Cognitive Thrashing)
  3. Efficiency (Time-to-Solve & Inference Cost)

Result 1: The Reasoning Ceiling

- **The "Knee" of the Curve:** All tested models achieve 85-95% of maximum performance by just 16k tokens.
- **Diminishing Returns:** DeepSeek-V3.2 and GPT-5.1 show 0.0% performance gain when scaling from 32k to 64k tokens.
- **Functional Equivalence:** A constrained 32k agent is statistically indistinguishable from an unconstrained 64k agent in solve rate, proving that architectural capacity ≠ reasoning capacity.

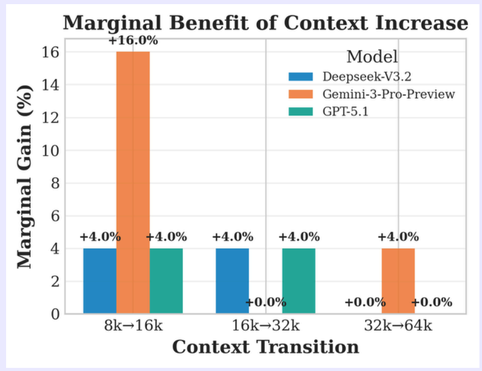


Figure 1: Marginal gain per context transition

Result 2: Anatomy of Failure

- **Context Starvation (8k):** High Hallucination Rate (3.2%); models fabricate files/ tools when history is truncated.
- **Context Overload (64k):** Shift to Cognitive Thrashing. Agents enter "Fragile Execution" loops, repeating failed commands.
- **ReAct Redundancy:** At high context, history is dominated by error traces. The agent biases towards prior failures rather than current objectives.

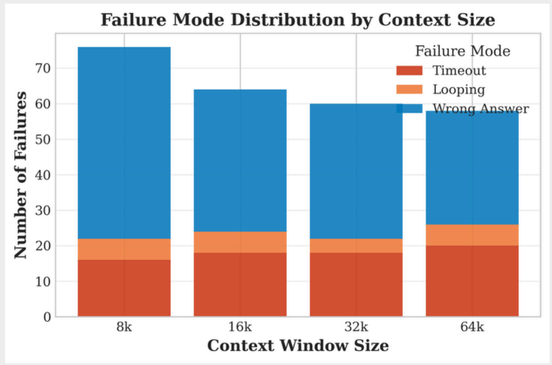


Figure 2: Failure mode distribution by context size. Wrong answers decrease with context while timeouts increase slightly

Discussion

- **Effective vs. Available Context:** The saturation at 16k proves that large windows act as interference, not memory. The model fails to prioritize signals over accumulated noise.
- **ReAct Redundancy:** At high context (64k), the agent's history is dominated by error traces. The model overfits to its own past failures, biasing it toward Cognitive Thrashing rather than fresh strategies.
- **The Economic Reality:**

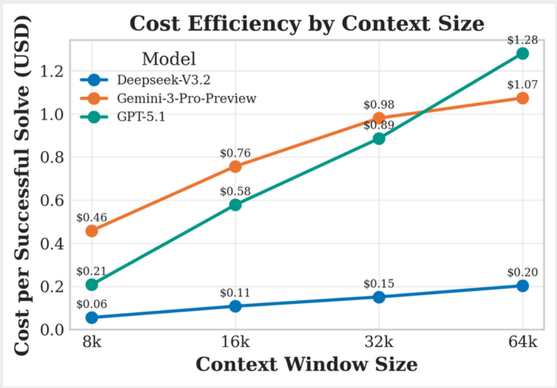


Figure 3: Cost per successful solve by context size. Diminishing returns are evident beyond 16k for all models.

- **Optimal Zone:** 16k–32k tokens represents the efficiency frontier.
- **The Waste:** Gemini-3-Pro-Preview at 64k costs 2x more ($21.47) than at 16k ($14.37) for a negligible 4% gain.

Next Steps

- **Compute-Aware Prompting:** Calibrate agents to recognize runtime complexity (e.g., avoiding brute-force on large keys) to reduce timeouts.
- **Smart Truncation:** Replace FIFO with Semantic Compression or RAG to retain critical signals while reducing noise.
- **Pass@k Evaluation:** Move from single-pass to multi-pass evaluation to account for non-deterministic sampling variance.

References

[1] Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan Rossi, Seunghyun Yoon, and Hinrich Sch¨utze. Nolima: Long-context evaluation beyond literal matching, 2025.
[2] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12:157–173, 2024.