

Context Compression for Long-Running Agents

A one-page operating checklist: budgets, boundaries, tiers, and the cost math.

AI Agents · Context Engineering · LLM · Companion to <https://aabee.tech/blog/context-compression-strategies-long-running-agents/> · October 2026

Before you compress anything

- ☐ Classify the session: retrieval / analysis / creative task
- ☐ Log per-turn tokens by source (system, context, tools, conversation)
- ☐ Log the cache-hit ratio per turn

Implementation order

1. Tool-output budgets at the tool layer (previews + file references)
2. Define task boundaries: checkpoint state, append-only between them
3. Tiered compression at boundaries: tool payloads → superseded reasoning → closed topics
4. Never compress the architectural-decisions layer
5. Canary recall check after each compression; roll back on 2 failures

Cost sanity math

- Mid-session summarization = full cache rebuild on every later turn
- Boundary compression amortizes the rebuild over the next stretch
- Aggressive compression can *raise* total cost despite fewer tokens

Rule of thumb: if the window is full of bulky intermediates, delegate to a sub-agent instead of compressing.

aabee.tech — AI agent engineering and software performance engineering, delivered as measurable outcomes. IT Park Uzbekistan Resident.

Questions or a project where this applies? b2b@aabee.tech · aabee.tech