ZENTHOS

ENGINEERING CHECKLIST

Context Engineering: 7 checks before your agent runs for hours

Long-context studies (Lost in the Middle 2023, Chroma's Context Rot 2025, NoLiMa 2025) all find the same thing: the 10,000th token is not processed as reliably as the 100th. Engineer for it.

- [ ] **Context budget set - and it's well below the advertised window**
  Cap working context (a common operating point: 40-60% of the window); count tokens per component.
- [ ] **Compaction checkpoints, not cliff-edge truncation**
  Summarize old turns into a structured note (decisions, open tasks, constraints, IDs); keep recent turns verbatim.
- [ ] **Durable facts live in external memory, not the transcript**
  Files or a DB the agent re-reads on demand; a memory it can grep beats a transcript it must attend over.
- [ ] **Sub-agents return conclusions, not transcripts**
  Fan work out to fresh-context workers; the orchestrator's context stays small.
- [ ] **Tool results pruned before they enter context**
  Raw API responses and file dumps are the biggest rot accelerant; store payloads out-of-band.
- [ ] **Instructions at the start, the live question restated at the end**
  Never bury the ask mid-context - middle positions are recalled worst.
- [ ] **A long-session eval exists and runs on every change**
  Replay representative sessions; track quality vs context length; alert on drift.

All seven checked? Your agent will still rot eventually - but predictably, measurably, and much later.