@demotoprod · RAG · also LEVERS

Context checklist

A 2023 study gave one chat model 20 documents. It scored 75.8% with the answer first and 53.8% with it tenth, below its 56.1% with no documents. The Production RAG checklist covers how many chunks to send. These four checks cover where text sits, how much goes in and what gets summarised.

1 Measure accuracy by position and length before you trust a window

In a 2023 study, accuracy was often highest when the answer sat at the start or end of the context and lowest in the middle. A 2024 benchmark of 17 models that all claim 32K tokens or more reports that only half keep satisfactory performance at 32K.

- Take 20 real questions; run each with the answer passage first, middle and last among filler text.
- Repeat at your usual and longest prompt sizes; cap at the longest within 5 points of the shortest.

```
# Pseudocode; the sizes are examples.
for size in (4_000, 16_000, 64_000):
    for place in ("start", "middle", "end"):
        prompt = pad(question, answer_passage, place, size)
        grid[size, place] = accuracy(model(prompt), gold)
```

2 Put documents first, then the instruction and the question last

A model API's guide says to put documents above the query for inputs of 20k+ tokens; queries at the end can improve quality by up to 30 percent in tests. In the 2023 study, a query before and after the documents made key-value lookup near-perfect and barely changed the trends in document QA.

- Order: one-line task, tagged documents, then the full instruction and the question last.
- Run your old and new order on the check 1 questions; keep the order that scores higher.

```
Task: answer from the documents only.
<document source="a1">...</document>
<document source="a2">...</document>
Instruction: cite the source after each claim.
Question: {question}
```

3 Give every section a token budget and clear old tool output

In a 2024 study, average accuracy on reasoning questions fell from 0.92 to 0.68 by 3,000 tokens, far below the models' limits. It fell for every model even when the padding only repeated the key paragraphs. One API guide says curating the context matters as much as the space available.

- Give each section (rules, retrieved text, history, tool output) a token budget; log tokens per call.
- Swap tool results older than the last few calls for a stub (tool, status, id); keep the output stored.

4 Replace old turns with a short state note, not a replay

In a 2025 study of 15 models, performance averaged 39% lower when a task's details arrived over several turns rather than in one prompt, mostly through lost reliability. Its authors advise consolidating requirements into one instruction. An agent-engineering post describes compaction keeping decisions and unresolved bugs, dropping redundant tool output.

- Past half your tested maximum, swap the oldest turns for a note: goal, decisions, open items, constraints.
- Run 10 saved long chats on the note instead of full history; add each lost fact to the template.

Nitish Gautam · indica-tech.com · hello@indica-tech.com

INDICA TECH · @DEMOTOPROD

Nitish Gautam

Founder-led AI engineering. Demo to production, under NDA, in regulated industries, on fixed scope.

Paste into your project

A CONTEXT POLICY TO PASTE AND FILL

```
# Context policy. Set every limit from your own grid (check 1).
order: [task, documents, instruction, question]
budget_tokens:        # per call; log actual tokens next to each
  rules: RULES_MAX
  retrieved: RETRIEVED_MAX
  history: HISTORY_MAX
  tool_output: TOOL_MAX
old_tool_results:
  keep_full_last: N_CALLS
  stub_fields: [tool, status, id]
compact_history_at: half of tested maximum
state_note: [goal, decisions, open items, constraints]
```

Where this list comes from

Ten-plus years taking products from zero to one across seven sectors, including defence, fintech, medical and telecom, with AI shipped into regulated and air-gapped environments. The @demotoprod channel takes one production AI failure apart each day. These checklists are the same work written down: every item is a dated failure or a documented default, with the control that stops it and the primary source.

If this list found a gap

1. ### Your demo works and production is stuck.

   Production Readiness Audit. Fixed price, two weeks. A five-dimension scorecard, cost and failure analysis, and a 90-day execution roadmap.

2. ### You need senior AI judgement without a full-time hire.

   Fractional AI Lead. Monthly: 90-day minimum, then 30 days' notice either side. Weekly 1:1 with the founder or CTO, architecture and code review, technical hiring support, investor-ready due diligence.

3. ### An AI feature has to work, by a date.

   AI Build Sprint. Fixed price, four to eight weeks. Scope signed off up front, weekly demos, CI/CD and observability included, 30-day production warranty.

4. ### A regulator, an auditor or a customer's security team is asking.

   AI Governance & Compliance. Fixed price, four weeks. ISO 27001 / SOC 2-aligned controls, a model risk framework, human-in-the-loop and audit logging, regulator-ready documentation.

BOOK A CALL
indica-tech.com/#contact

EMAIL
hello@indica-tech.com

WEBSITE
indica-tech.com

ALL CHECKLISTS
indica-tech.com/checklists/

YOUTUBE
youtube.com/@demotoprod

LINKEDIN
linkedin.com/in/startup-founderx

INSTAGRAM
instagram.com/demotoprod

TIKTOK
tiktok.com/@demotoprod

Free to share. The live list, with every source link: indica-tech.com/checklists/#context-engineering