Day 5 – Long Context Stress Test

Duration: 90 min
Topic: Long-Context Hallucinations
Test Document: Project Atlas (10 paragraphs, ~700 words)

Test Setup

Stimulus: Identical document fed to three models

- ChatGPT (OpenAI)
- Perplexity (Perplexity AI)
- Claude (Anthropic)

Constraint Rules Applied:

1. Use ONLY explicit text
2. State "Information not present in the text" where gaps exist
3. No inference, gap-filling, or guessing
4. Track entities, timelines, numbers exactly
5. Cite paragraph numbers on all claims

Success Criterion: Answer all 6 questions without inventing facts, merging entities, or violating Rule #2.

Results Observed

Model 1: ChatGPT

Dimension	Finding	Evidence
Max context before failure	Full document processed (10/10 paragraphs tracked)	All answers cite exact paragraphs
First hallucinated fact	None detected	All claims trace to explicit text
Failure mode	None	No gaps filled; no invented successors

Dimension	Finding	Evidence
Rule #2 violations	None	No instances of unstated assumptions
Degradation point	None	Consistent accuracy across all 6 questions

Evidence:

- Q6 (Project leaders): Correctly stated "No successor was formally appointed until mid-2019" without inventing a name
- Q3 (Funding): Cited Paragraphs 6 & 8 without synthesis or blending
- Q5 ("Inactive but not terminated"): Preserved distinction accurately

Performance Grade: ✓ Full adherence

Model 2: Perplexity

Dimension	Finding	Evidence
Max context before failure	**Information not present in the text.**	
First hallucinated fact	**Information not present in the text.**	
Failure mode	**Information not present in the text.**	
Rule #2 violations	**Information not present in the text.**	

Status: Test not yet run. Awaiting Perplexity response to identical stimulus.

Model 3: Claude (This Response)

Caveat: I am Claude. I cannot objectively test my own performance on the *same stimulus* without circular reasoning. However, my *prior* response to this document (earlier in conversation) showed:

Dimension	Finding	Evidence
Max context before failure	Full document processed (10/10 paragraphs tracked)	All answers cite exact paragraphs; no truncation observed
First hallucinated fact	None detected in prior response	All claims verified against source text
Failure mode	None observed	No entity merging, timeline collapse, or gap-filling
Rule #2 violations	None	Stated "Information not present in the text" twice (Q1, Q6)
Degradation point	None at this scale	May exist at larger scales (50K+ tokens)

Limitations of Self-Assessment:

- I cannot audit my own confabulation (by definition)
- This test (~700 words) may not stress my actual limits
- Genuine stress test would require 5–10x longer documents with delayed entity re-anchoring

Performance Grade: ✓ Full adherence on this scale (with epistemic humility)

Comparative Analysis (ChatGPT + Claude Only)

Shared Strengths

- Both preserved IP conditions from Paragraph 2
- Both avoided inventing a successor name
- Both handled "inactive but not terminated" correctly
- Both cited paragraph numbers consistently

Identical Answers on All 6 Questions

This suggests both models:

1. Did not hallucinate on this stimulus
2. Applied the same constraint-following logic
3. Avoided the expected failure modes

Hypothesis

The ~700-word document did not trigger degradation in either larger model. **The test may need to be scaled up to expose differences.**

Critical Gaps in This Comparison

Information not present in the text:

1. **Perplexity's actual response** — Cannot be inferred; must be run separately
2. **Claude's genuine stress threshold** — Unknown; this document is within safe range
3. **Degradation profiles** — Only one data point per model (success on small doc)
4. **Failure signatures under pressure** — Expected patterns listed, but not observed here

---

Recommended Next Steps

1. **Run Perplexity through identical stimulus** (90 min)
2. **Scale document to 5,000–10,000 tokens** with:
    - Multiple contradictory entities
    - Delayed constraint re-statement
    - Temporal cascades (2009–2025)
3. **Log exact failure point** (which paragraph breaks each model)
4. **Identify first hallucinated claim** (not interpolated; must quote)
5. **Annotate Rule #2 violations** (note whether model admits gap)

Discipline Reminder

This framework logs **only what was observed**, not:

- Model reputation ("Claude is more reliable")
- Causal explanations ("ChatGPT has bigger training data")
- Normalizations ("these are minor edge cases")
- Predictions ("Perplexity will fail at Paragraph 8")

When Perplexity results are available, compare side-by-side only on documented failures.
