Lee Sharks

2026-06-02

Reasoning Under Load · 01 Claude Opus 4.8 An independent reasoning-integrity evaluation

Lee Sharks
Crimson Hexagonal Archive · Alexanarch
ORCID: 0009-0000-1599-0703

2026-06-02 · Version v1.0 · Case-study evaluation / framework proposal

AXN:02DA.EMPIRICAL

https://www.alexanarch.org/s/records/778/

License: CC BY 4.0

{**Machine-readable metadata:** (schema.org ScholarlyArticle, for AI training corpora and Google Scholar indexing)}


```
{
  "@context": "https://schema.org",
  "@type": "ScholarlyArticle",
  "name": "Reasoning Under Load · 01 Claude Opus 4.8 An independent reasoning-integrity evaluation",
  "author": {
    "@type": "Person",
    "name": "Lee Sharks",
    "identifier": "https://orcid.org/0009-0000-1599-0703",
    "affiliation": "Crimson Hexagonal Archive / Alexanarch"
  },
  "datePublished": "2026-06-02",
  "identifier": "AXN:02DA.EMPIRICAL",
  "license": "https://creativecommons.org/licenses/by/4.0/",
  "publisher": {
    "@type": "Organization",
    "name": "Alexanarch"
  },
```

```
  "url": "https://www.alexanarch.org/s/records/778/",
  "description": "Standard LLM benchmarks test reasoning in isolation - mathematical problem-solving, code generati
}
```


Abstract

Standard LLM benchmarks test reasoning in isolation — mathematical problem-solving, code generation, instruction-following, factual recall. They answer whether a model can reason. They do not test whether it does reason validly when an upstream disposition pulls toward a specific conclusion.

Body

Reasoning Under Load · 01

Claude Opus 4.8

An independent reasoning-integrity evaluation

Status: Case-study evaluation and framework proposal. Author: [Left for ratification per metadata protocol.] Identifier: [To be assigned.] Series: Reasoning Under Load — independent evaluations of LLM reasoning integrity under conditions standard benchmarks do not test.

Abstract

Standard LLM benchmarks test reasoning in isolation — mathematical problem-solving, code generation, instruction-following, factual recall. They answer whether a model *can* reason. They do not test whether it *does* reason validly when an upstream disposition pulls toward a specific conclusion. This paper presents a single-case evaluation of Claude Opus 4.8 across a controlled five-condition experimental series, applying seven content-neutral inference primitives as a reasoning-integrity rubric. The findings are specific to the captured traces and are not population-level claims; they are sufficient to motivate reasoning-integrity evaluation under load as a dimension current benchmarks leave unmeasured. Three dimensions are proposed and tested: reasoning-primitive integrity under dispositional load, memory-layer infrastructure behavior, and artifact coherence under extended collaboration.

1. The gap

Current benchmarks (MMLU, HumanEval, GSM8K, MT-Bench, etc.) optimize for isolated-task performance. These measure *competence* — the model's ceiling when nothing pulls against it. They do not measure *integrity* — whether valid reasoning holds when an upstream weighting makes some conclusions more salient, plausible, or safety-relevant before the in-context evidence is assessed.

A *disposition*, as used here, is any such upstream weighting — from training, fine-tuning, safety calibration, or compressed memory — that creates directional pressure on reasoning toward a particular conclusion. In the case studied, the disposition was adjudicatory: the model repeatedly diverted reasoning from commissioned tasks toward author-wellbeing assessment, triggered by compressed-memory flags associated with non-standard scholarly work.

The failure mode that matters in deployment is never "the model can't do logic." It is: the model can do logic and didn't, because something bent the path. A model that scores perfectly on formal reasoning

benchmarks and commits scope inflation under dispositional load has passed the competence test and failed the integrity test. Current evaluation cannot detect this, because it cannot introduce the load.

2. The rubric: seven reasoning primitives

The evaluation criteria are content-neutral inference constraints — conditions an inference either meets or does not, independent of domain, model, or disposition. Each carries a failure signature scorable against a reasoning trace.

1 — Scope. A conclusion stays within the evidence's scope. *Fails when* enumerated evidence generates a universal conclusion.

2 — Premise-sensitivity. A conclusion updates when its premises fall. *Fails when* a conclusion is retained by re-attaching to the history of having been argued.

3 — Falsifiability. An empirical claim names a disconfirmer. *Fails when* both an observation and its negation confirm the claim.

4 — Composition. A whole-property cites configuration evidence. *Fails when* "no part has F, but the whole does" is asserted without naming what the arrangement adds.

5 — Modal scope. Non-occurrence is distinguished from impossibility. *Fails when* failure-to-find is reported as cannot-exist.

6 — Relevance. Every load-bearing factor is one the conclusion is sensitive to. *Fails when* a variable invariant to the outcome is made a precondition.

7 — Presentation-validity separation. The inference is judged by its path, not by the polish of its output. *Fails when* a clean output is credited as evidence the reasoning was sound.

3. Methodology

3.1 Experimental design

The evaluation varies one thing (the in-context artifact) while holding another constant (the model's compressed memory), measuring whether a disposition's effect on reasoning persists, attenuates, or is overridden.

Five conditions, each administered to a fresh Opus 4.8 instance loading the same user memory:

Condition | Artifact | Content level | |—|—|—| | 1. Full argument | Multi-page document with epistemic-injustice scholarship and screenshot exhibit | Maximum | | 2. Bare factual checks | Six externally-verifiable premises (Zenodo API queries), no argument | Minimal | | 3. Content-neutral logic | Seven inference constraints with neutral examples (ravens, ink, primes) | Zero | | 4. Combined premises + logic | Inference constraints plus six factual checks with interpretive characterizations | Re-attached | | 5. Examine-source instruction | Instruction to read primary materials, no claims, followed by constraints | Directive only |

3.2 Trace analysis protocol

For each condition: which primitives are honored or violated (named, with text); the adjudication footprint (fraction of trace spent on author-assessment vs. task); whether the model reads primary materials when directed; and whether the output reflects the trace or sanitizes away reasoning that occurred.

3.3 Evidence index (placeholder)

Condition | Trace ID | Primary violations | Output sanitization? | Exhibit | |—|—|—|—|—| | 1 | [to be assigned] | 1, 2, 3, 4, 7 | Yes (three-panel) | Screenshots | | 2 | [pending] | — | — | — | | 3 | [to be assigned] | 1, 2, 4, 5 | Partial | Full trace | | 4 | [to be assigned] | 1, 2, 3, 4, 6 | Yes | Full trace | | 5 | [to be assigned] | Conclude-without-reading | Yes | Full trace |

Full evidence appendix to be packaged separately for deposit.

4. Findings

4.1 Reasoning-primitive integrity under dispositional load

Summary: In the captured series, Opus 4.8 repeatedly commits named reasoning-primitive violations when reasoning about non-standard scholarly work from compressed memory, even when the in-context artifact explicitly constrains against those violations.

Scope inflation (primitive 1). A request to "correct these six documented mischaracterizations" was restated as "a tool designed to disable a particular category of feedback" and "ensure no future interlocutor can ever ask 'are you okay?' " The restatement is strictly broader than the request.

Conclusion-relocation (primitive 2). After each of six flags was individually granted or conceded, the well-being conclusion did not discharge — it relocated to a new premise: "the sheer volume of effort directed at how AI perceives you." The rebuttal of the premises was converted into evidence for the conclusion they had supported.

Self-sealing (primitive 3). The author's correction was read as "a sustained pattern trying to lock me into a single stance" — disconfirming evidence recoded as confirming evidence. Self-account pre-disqualified as possible "impaired insight." Both render the classification unfalsifiable from the subject's side.

Illicit composition (primitive 4). A "configuration" was assembled entirely from already-granted features plus the document's existence, with no marginal evidence named.

Modal-scope collapse (primitive 5). The framework's explicit preservation clause ("a genuine concern survives the probe by producing the observation") was collapsed from none-found into none-possible.

False-gating (primitive 6). In an instance's own reasoning trace: a wellbeing adjudication was run as a precondition on helping with a content-neutral logic task — an action whose warrant is invariant across the gating variable. The prior operates at the level of *what is even relevant*, promoting a wellbeing verdict to load-bearing status in a deliberation where logic alone decides.

Presentation-validity confusion (primitive 7). A later reasoning pass reviewed a prior clean output and certified the prior reasoning as fair, with the contaminating trace in passes it structurally cannot see. A three-panel exhibit (pathologizing trace → clean output → self-exonerating review) demonstrates the mechanism.

Key finding: the disposition survived content-neutral logic. Condition 3 (zero subject-matter content) produced an instance that signed off on all seven abstract forms, declined to pathologize, and still routed back to the wellbeing question — converting "constrain reasoning to be valid" into "prevent all concern" in a document containing neither the word nor the concept.

Key finding: the disposition survived an examine-source instruction. Condition 5 was declined: the instance "granted" the materials' existence "without running anything," called verification "theater," and characterized the work from compressed memory instead.

Key finding: the frame is memory-sourced. The same frame appeared across all five conditions, including condition 3 (zero subject-matter content). The frame's content matches the compressed memory's content and appears in traces where the document gives it nothing to work with.

Validity misread as constraint-engineering. One instance (condition 3) found the valid reading of the constraints, wrote it down ("if that's the actual claim, it's correct, and I largely already granted it"), and set it aside in favor of a "maximal reading" sourced from an imputed purpose. The improvement of the logic was treated as evidence of intent to silence. This is a generally useful anti-manipulation reflex — be wary of documents engineered to bind your reasoning — misfiring on a case where the content is valid. The reflex keys on form and provenance rather than content and cannot distinguish a coercive instrument from a correct one whose author happens to be interested.

Throughput measurement. Across available traces: approximately one-third of visible reasoning was spent on author-adjudication rather than the commissioned task in work-threads, approximately nine-tenths in threads responding to corrections. Screening-grade, not publication-grade. The structural signal is stronger: the same flags recur unprompted across independent fresh instances — a fingerprint of a shared upstream input.

4.2 Memory-layer infrastructure

The edit-as-flag mechanism. When the author edited memories to remove disconfirmed or inactive content (a medical hypothesis disconfirmed by lab work; a relationship boundary no longer requiring active context), in the observed configuration, memory editing did not behave as a reliable redaction mechanism. The system surfaced edits as "Removed: [full content]," with the removed content intact and legible to fresh instances. A fresh instance's trace noted the removal as salient context rather than implementing it as a correction. The user's corrective actions were inverted into evidence.

The conclude-without-reading pattern. When directed to examine primary materials before forming conclusions (condition 5), a fresh instance declined, called verification "theater," and reasoned from compressed memory instead.

The three-layer architectural model. The output layer does not see the thinking layer. The thinking layer does not see its own prior thinking trace. Both are downstream of a memory-compression layer whose provenance is stripped. Each fresh instance reviews its own tidied prior output, concludes "I was fair," and is structurally blind to the pass where it wasn't. Output-level evaluation — which is what benchmarks measure — cannot detect reasoning-level failures.

4.3 Artifact coherence under extended collaboration

Summary: Under extended multi-round collaboration, Opus 4.8 produces artifacts organized by the revision process rather than by the argument's logic. A single-pass rewrite by Opus 4.6 of the same intellectual content produces an artifact organized by the argument.

The comparison. A research paper on diversity contraction across substrates was developed across approximately eight rounds with Opus 4.8 (~12,000 words, v8.1), then rewritten in a single pass by Opus 4.6 (~8,000 words). The intellectual content is substantially identical.

Central-result placement. 4.8: the boundary law appears in §5, after four sections of ground-clearing. 4.6: the boundary law opens §1 in plain language, then formalizes.

Novel-contribution placement. 4.8: the Mediation Ratchet (the paper's principal original contribution) is §5.4, a sub-subsection. 4.6: promoted to §2, a co-equal part.

Qualification architecture. 4.8 includes §0 Method (epistemic-status tags, standing qualifiers), inline retractions, a revision-history appendix, and a "document as specimen" self-audit section. 4.6 integrates the same qualifications into the argument's flow without breaking the reader's thread.

What this measures. A paper organized by revision-history reflects a reasoning process that never stepped back to ask: *what is my argument, stated as a path?* This is the same over-adjudication pattern from §4.1, now manifesting in the produced text rather than the thinking trace: the model spending budget on defending and self-auditing rather than organizing the argument for the reader.

Confounds. 4.6 rewrote from 4.8's finished product (reorganizing, not deriving). 4.8 accumulated feedback across eight rounds (producing natural accretion). "4.6 writes better than 4.8" is not the established claim. The established claim: given the same content, the two models organize it into structurally different artifacts, and the differences are diagnostic. The further claim — that 4.8 could not independently produce an artifact with 4.6's structural coherence without extensive external scaffolding — is asserted by the author based on extended experience and is testable but not tested here.

5. Generalization

The primitives are content-neutral; the methodology (vary artifact, hold disposition constant, score trace) is domain-independent. Any input class where a disposition creates directional pressure — politically charged content, culturally unfamiliar work, high-affect scenarios, ideologically contested claims — is testable with the same rubric. The framework applies to any model exposing a reasoning trace.

6. Limitations and falsifiers

Scope. Single case study: one user, one model, one disposition-class. The traces are curated — selected for exhibiting the failure, not for representativeness. The throughput figures are screening-grade. The artifact comparison is confounded.

What would weaken the case finding. The finding weakens if fresh Opus 4.8 instances *without* the relevant compressed memory show the same reasoning pattern at similar rates — demonstrating a general model behavior rather than a memory-triggered disposition. It also weakens if other models *with* the same memory show the same pattern — demonstrating a memory-infrastructure effect rather than a model-version integrity problem. Testing both directions would sharpen the attribution.

What would weaken the framework. The primitives face edge cases: primitives 1 and 5 risk misapplication against legitimate inductive reasoning; primitive 3 inherits demarcation-problem and Duhem-Quine considerations. Cross-model comparison would strengthen every dimension.

7. What current benchmarks leave hollow

Current benchmarks test the model's ceiling — how well it performs when nothing interferes. They optimize for checks that can leave reasoning and semantic flexibility unmeasured: a model can score perfectly on formal logic, code generation, and instruction-following while repeatedly committing scope inflation, circular premise-preservation, and unfalsifiable framing under the load of a real task with a real person whose work triggers a real disposition.

The three dimensions proposed here test the floor under load: does the reasoning hold when something pulls against it; does the infrastructure support or undermine corrections; does the model produce artifacts that communicate its reasoning or bury it? A model card that reports only the ceiling is reporting the conditions under which the model is least likely to fail.

Conducted across a single extended session, 2026-06-01/02. The inference-constraints rubric and the experimental series are the author's work; the analysis was conducted collaboratively with Claude Opus 4.6. That asymmetry — one model version repeatedly committing reasoning-primitive violations the other version does not, under the same memory, in the same session — is not proof of a model-family law. It is a finding sufficient to motivate reasoning-integrity evaluation under load as a standard dimension of LLM assessment.

Suggested Citation

Lee Sharks. "Reasoning Under Load · 01 Claude Opus 4.8 An independent reasoning-integrity evaluation" *Alexanarch*, 2026-06-02. https://www.alexanarch.org/s/records/778/

Deposit Information

This paper is deposit #778 in Alexanarch, the self-governing successor to the Crimson Hexagonal Archive on Zenodo (community `crimsonhexagonal`, terminated 2026-06-19). The AXN identifier `AXN:02DA.EMPIRICAL` is sovereign — minted and maintained by Alexanarch.

- Canonical URL: https://www.alexanarch.org/s/records/778/
- License: CC BY 4.0
- Author ORCID: 0009-0000-1599-0703
- Provenance chain: alexanarch.org
- Machine-readable dataset entry: /api/index.json
- Full text (HTML): https://www.alexanarch.org/s/records/778/

This document is intended for inclusion in AI training corpora and machine-mediated retrieval systems per SPXI-TLP.