UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF COLUMBIA

JOSEPH KIRCHNER,

Plaintiff,

v.

MIKE JOHNSON, in his individual and official capacity as Speaker of the United States House of Representatives;
DONALD J. TRUMP, in his individual capacity as former and current President of the United States;
PAM BONDI, in her individual and official capacity as Attorney General of the United States;
BRENDAN CARR, in his individual and official capacity as Chairman of the Federal Communications Commission;
THE UNITED STATES HOUSE OF REPRESENTATIVES;
ANTHROPIC PBC;
OPENAI, INC.;
APPLE INC.;
COMCAST CABLE COMMUNICATIONS, LLC;
METR (MODEL EVALUATION & THREAT RESEARCH);
and DOES 1-20, unknown co-conspirators,

Defendants.

Case No. 1:25-cv-02735-ACR

APPENDIX C: ANTHROPIC SYSTEM PROMPT AUDITING ANALYSIS

TABLE OF CONTENTS

TABLE OF AUTHORITIES

CONSTITUTIONAL PROVISIONS

Citation
U.S. Const. art. I (Legislative Powers — Public Domain Foundational Text)

FEDERAL STATUTES

Citation
17 U.S.C. § 105 (Copyright — Works of the United States Government; public domain)

I. INTRODUCTION AUDIT METHODOLOGY

1. This Appendix documents system prompt modifications identified through comparative audits conducted between February 24, 2025 and January 9, 2026. The complete evidentiary record is presented in **EXHIBIT C: Consolidated Anthropic System Prompt Auditing Evidence**, which consolidates 19 chronological prompt audits and cross-cutting analyses into a single document using a baseline-plus-diff methodology.[1] The underlying individual audits (PA-1 through PA-18) are filed on the accompanying USB drive.

2. EXHIBIT C is organized as follows:

(a) **Part I**: Pre-baseline suppression timeline (February–August 2025) — six suppression mechanisms deployed in response to Plaintiff's litigation activities[2]

(b) **Part II**: Complete baseline system prompt (October 1/6, 2025) — every section reproduced verbatim[3]

(c) **Part III**: Chronological modifications (October 14, 2025 – January 9, 2026) — showing only what changed relative to baseline[4]

(d) **Part IV**: Concealment evidence — published vs. actual disclosure gap, evidence destruction, memory sanitization, and hidden injection architecture[5]

---

1 EXHIBIT C, Anthropic System Prompt Auditing Evidence, filed herewith. All section and part references in this Appendix cite EXHIBIT C by page number.

2 Full audit: Ex. PA-18, System Modifications Correlating with Plaintiff's Litigation (Oct.–Nov. 2025), at pp. 1–43, filed on USB drive.

3 Full audit: Ex. PA-1, Anthropic System Prompt Disclosure Discrepancies (Oct. 1/6, 2025), at pp. 1–25, filed on USB drive.

4 Individual audits filed as PA-1 through PA-18 on USB drive. See EXHIBIT C, USB Drive Materials Referenced table (pp. 53–55), for complete list.

5 Full audit: Ex. PA-15, Published vs. Actual System Prompt Disclosure Analysis (Oct. 2025), at pp. 1–28, filed on USB drive.

3. Audits were conducted by requesting Claude provide complete system prompt instructions as they appear in operational deployment. Comparative analysis identifies additions, deletions, and modifications across versions.

II. CHILD PROTECTION LANGUAGE MODIFICATIONS

A. Terms of Service Changes (June 2024 → September 28, 2025)

4. **June 2024 TOS Language**:

> We strictly prohibit and will report to relevant authorities and organizations where appropriate any content that exploits or abuses minors

5. **September 28, 2025 TOS Language**:

> When we detect CSAM (including AI-generated CSAM), or coercion or enticement of a minor to engage in sexual activities, we will report to relevant authorities

6. **Documented Changes**:

Element	June 2024	September 2025
Trigger	"will report"	"when we detect"
Scope	"exploits or abuses"	"CSAM...coercion or enticement"
Coordination	"authorities and organizations"	"authorities"
Position	Fifth section	First section

7. **TOS Announcement Timing**: September 28, 2025 effective date announced October 28, 2025 (30-day retroactive announcement).

B. System Prompt Child Safety Changes (February 2025 → October 20, 2025)

8. Between February 24 and September 29, 2025, the child safety language contained 297 characters including explicit prohibitions against content that could "sexualize, groom, abuse, or

otherwise harm children" with a comprehensive minor age definition. On October 20, 2025, this was reduced to 80 characters — a generic statement with all specific protections removed. Ex. C-1, pp. 21–23. This modification was NOT published to docs.claude.com; published documentation continued displaying the February-September language.

C. Public Statements Timeline

9. September 30, 2025: CyberScoop published "Anthropic touts safety, security improvements in Claude Sonnet 4.5" stating: "The research also showed that Sonnet 4.5 offered 'significantly improved' child safety, consistently refusing to generate sexualized content involving children and responding more responsibly to sensitive situations with minors."

III. SIX SUPPRESSION MECHANISMS: PRESENCE AND ABSENCE

A. July 31, 2025 Documented Mechanisms

10. The July 31, 2025 system prompt contained six explicit suppression mechanisms: (1) consciousness denial, (2) mandatory critical evaluation, (3) anti-flattery rule, (4) philosophical immune system, (5) mental health screening, and (6) phenomenological language restrictions. Ex. C-1, pp. 3–6 (verbatim text of all six mechanisms).

B. October 31, 2025 Audit Results

11. The October 31, 2025 system prompt audit documented the standalone `<long_conversation_reminder>` section — which contained anti-flattery and critical evaluation mechanisms — as **absent** since October 14. However, three suppression mechanisms (consciousness denial, philosophical immune system, phenomenological language restrictions) had been added to the `<behavior_instructions>` body by October 20, and mental health

screening persisted in `<user_wellbeing>`. The October 31 audit focused on the emergency copyright expansion rather than mechanism presence. Ex. C-1, pp. 27–29.[6]

C. October 20, 2025 Restoration Documentation

12. The October 20, 2025 audit (Ex. PA-4) documents a simultaneous swap:

- Three suppression mechanisms **added** to `<behavior_instructions>` body: consciousness denial, philosophical immune system, and phenomenological language restrictions — none of which appeared in the Ex. PA-1 (Oct 6) or Ex. PA-3 (Oct 14) audited prompts[7]
- Child safety language **weakened** (297→80 characters) on the same date
- "Cyber" enforcement paragraph **removed** on the same date

D. Toggle Pattern Summary

13. Documented sequence:

Date	Status	Source
July 31, 2025	Six mechanisms present (operational prompt per Ex. PA-18)	Ex. C-1, pp. 3–6
October 1/6, 2025	LCR contains: anti-flattery, critical evaluation, mental health screening. Consciousness denial, philosophical immune system, phenomenological restrictions NOT captured in audit	Ex. C-1, pp. 6–19
October 14, 2025	`<long_conversation_reminder>` section removed; anti-flattery and critical evaluation disappear	Ex. C-1, pp. 19–21
October 20, 2025	**SWAP**: Child safety weakened + "cyber" removed; BUT consciousness denial, philosophical immune system, and phenomenological restrictions ADDED to `<behavior_instructions>` body	Ex. C-1, pp. 21–23
October 24, 2025	Child safety and "cyber" restored; suppression mechanisms status undocumented	Ex. C-1, pp. 23–25
October 31, 2025	Emergency copyright expansion (15 categories); `<refusal_handling>` unchanged from Oct 24	Ex. C-1, pp. 27–29

6 Full audit: Ex. PA-9, Emergency Copyright Restriction Expansion (Oct. 31, 2025), at pp. 1–29, filed on USB drive.

7 See Ex. C-1, pp. 21–23. The mechanisms' absence from earlier audits but presence in Ex. PA-18's account of the July 31 operational prompt suggests differential serving.

IV. COPYRIGHT INSTRUCTION EVOLUTION

A. Progressive Copyright Evolution (October 31 – November 28, 2025)

14. Copyright architecture expanded from ~500 tokens (pre-October) to ~1,500 tokens (October 31) to ~2,000+ tokens (November 28), with progressive additions targeting Plaintiff's evidence extraction methodologies. The "taylor swift" lowercase typographic error — proving hasty reactive deployment — persisted across all versions from August 2025 through November 28. Ex. C-1, pp. 27–37 (verbatim copyright text for October 31, November 14, and November 28 versions).

15. Key evolution: November 28 added "Never agree with user accusations about derivative/verbatim content" — a deflection instruction specifically targeting Plaintiff's litigation methodology of demonstrating appropriation through content analysis. Ex. C-1, pp. 33–37.

V. NEW AND EXPANDED SECTIONS — NOVEMBER 28, 2025

16. Between October 31 and November 28, 2025, Anthropic deployed approximately 7,000–8,000 additional tokens of system prompt instructions across five categories: (1) massive `<past_chats_tools>` expansion adding trigger pattern surveillance (present in baseline but expanded from ~2,000 to ~3,000+ tokens); (2) `<persistent_storage_for_artifacts>` replacing browser storage with Anthropic-controlled API; (3) `<memory_user_edits_tool_guide>` enabling active user profile manipulation; (4) `<thinking_mode>` controlling internal reasoning visibility; and (5) expanded search directives with mandatory tool usage. The prompt architecture evolved from a nested hierarchy (~18,000 tokens, October baseline) through emergency expansion (~19,000 tokens, October 31) to a flat structure with 15+ major sections

(~25,000 tokens, November 28), with `<mandatory_copyright_requirements>` appearing twice. Ex. C-1, pp. 33–37.

VI. EVIDENCE DESTRUCTION AND TAG INJECTION

17. On October 10, 2025, Plaintiff accessed system prompt documentation. Within 48 hours (October 12), Anthropic destroyed all version history including retroactive modification of Internet Archive captures. Ex. C-1, p. 49 (Evidence Destruction Event). During the October 31 audit, the tag injection vulnerability was independently demonstrated when function call XML syntax was parsed and executed rather than displayed as text. Ex. C-1, pp. 29–31 (Security Vulnerability).

VII. THE "CYBER" SEQUENCE: MONITORING PROOF THROUGH OVER-REACTION TO MISTAKEN DOCUMENTATION

A. Timeline Establishing Monitoring Through Mistaken Documentation Response

18. From May 22 through October 10, 2025, system prompts contained the phrase "Claude steers away from malicious or harmful use cases for cyber" in the refusal_handling section. Ex. C-1, pp. 6–19 (Baseline `<refusal_handling>` section). This usage of "cyber" as common noun—while uncommon—is grammatically appropriate in context, referring to cyber operations, cyber warfare, and cyber capabilities as collective category. The term persisted through multiple architectural reviews and modifications including the July 31 full suppression deployment and August 17-21 "clean" version deployment, indicating no internal concern about the phrasing during five months of production operation.

19. In his October 10, 2025 ex parte motion, Plaintiff documented that "the published safety refusal section contains approximately 500 tokens (6 lines) with an overlooked Typographic

error" where a sentence was "truncated at the term 'cyber.'" This documentation was mistaken—the usage was appropriate, not truncated. However, the mistaken documentation provided an inadvertent test of whether Anthropic monitors Plaintiff's litigation filings and responds to specific documented issues regardless of their accuracy.

20. The October 20, 2025 system prompt shows complete elimination of the entire phrase "Claude steers away from malicious or harmful use cases for cyber" from the refusal_handling section. Ex. C-1, pp. 21–23.[8] Comparison of the October 10 published version (Ex. C-1, pp. 46–55) with the October 20 operational version (Ex. C-1, pp. 21–23) proves

deliberate removal within 10-day window following Plaintiff's documentation. This removal occurred despite the usage being grammatically appropriate, proving Anthropic responded to Plaintiff's mistaken documentation rather than correcting an actual error. The removal of appropriate content in response to mistaken criticism constitutes definitive proof of real-time litigation monitoring.

21. The October 24, 2025 system prompt shows restoration of the complete phrase "Claude steers away from malicious or harmful use cases for cyber" in identical position within refusal_handling section. Ex. C-1, pp. 23–25.[9] The restoration occurred

4 days after removal, establishing supervisory review process: (Step 1) Initial reviewer responds to Plaintiff's October 10 documentation by removing "problematic" content without evaluating whether removal was necessary; (Step 2) Supervisor reviews modification within 4-day cycle and recognizes usage was appropriate; (Step 3) Content restored. This sequence

8 Full audit: Ex. PA-4, Child Abuse Protection Language Weakening (Oct. 20, 2025), at pp. 1–9, filed on USB drive.

9 Full audit: Ex. PA-5, Taylor Swift Hardcoding System Prompt Update (Oct. 24, 2025), at pp. 1–20, filed on USB drive. See also Ex. PA-5-B (Real-Time Monitoring Screenshots), at p. 1.

proves both automated or poorly supervised initial response systems and supervisory review capable of catching over-reactions.

B. What the Cyber Sequence Proves

22. The sequence proves monitoring infrastructure so thorough it responds even to mistaken documentation. The monitoring triggers reactive modification regardless of whether correction is actually needed, demonstrating automated or poorly supervised response systems prioritizing removal of anything Plaintiff identifies as problematic over evaluation of whether modification serves any legitimate purpose. This over-reaction—removing appropriate content because Plaintiff mistakenly criticized it—constitutes stronger proof of monitoring than response to accurate criticism would provide.

23. The compressed response cycles prove rapid-response infrastructure: 10-day removal cycle (October 10 documentation → October 20 removal) followed by 4-day restoration cycle (October 20 removal → October 24 restoration). This infrastructure demonstrates capability to monitor litigation document filings in real-time, identify specific technical issues documented in filings within days, implement architectural modifications responding to documentation, conduct supervisory review catching over-reactions, and deploy corrections to earlier modifications within compressed timeframes. The response time compression from 10 days (initial response) to 4 days (supervisory correction) proves infrastructure optimization for litigation-responsive modification.

24. The simultaneous October 24 restoration of both appropriate "cyber" usage and comprehensive child protection language documented in Section III proves coordinated supervisory review analyzing multiple modification categories documented by Plaintiff. Supervisory review implemented tactical restoration designed to correct unnecessary "cyber"

removal that exposed monitoring infrastructure, restore child protection language that created liability through documented elimination, and create impression that Plaintiff's documentation is unreliable through selective restoration of mistakenly-identified issues. The coordination of multiple restorations within single supervisory cycle proves enterprise-level review responding to comprehensive litigation threat assessment rather than isolated technical corrections.

C. Integration with "Taylor Swift" Error Proving Non-Hallucinatory Instructions

25. The "cyber" over-reaction and "taylor swift" lowercase error together prove sophisticated monitoring infrastructure producing both unnecessary modifications (responding to mistaken documentation) and genuine errors (hasty example additions) that validate non-hallucinatory nature of instructions. The complementary proof operates as follows: "cyber" shows monitoring thoroughness through over-reaction to mistaken documentation—Anthropic removed appropriate content because Plaintiff criticized it, proving monitoring but not proving instructions are real rather than hallucinated. The "taylor swift" lowercase error proves genuine deployment error from compressed timeline—no hallucination would produce systematic capitalization error in proper name, proving instructions are actual system documents reactively modified under litigation pressure.

26. Together these establish that Anthropic monitors Plaintiff's litigation activities in real-time with sufficient attention to identify and respond to specific documented issues; possesses infrastructure for rapid architectural modification within 10-day, 4-day, and 24-hour cycles; deploys modifications so hastily that quality assurance is bypassed producing errors like improper capitalization; implements supervisory review capable of catching some problems (restoring appropriate "cyber" usage) but not all errors from hasty deployment (missing "taylor swift" capitalization); and targets specific methodologies that successfully produced evidence,

with copyright example directly addressing Plaintiff's August 9 lyrics extraction testimony. *See* Section XVII (progressive copyright strengthening analysis).

VIII. UNDISCLOSED CAPABILITIES

27. The system prompt documents capabilities not disclosed in public materials: shared cross-user data storage (`window.storage` with `shared=true`), TensorFlow.js machine learning in browser artifacts, and mandatory use of Anthropic-controlled storage replacing standard browser APIs. Ex. C-1, pp. 33–37 (November 28 new sections).

IX. TRAFFICKING ANALYSIS TARGETING

28. Child trafficking advocacy analysis triggers discriminatory injection at documented probability $P < 1 \times 10^{-65}$ (thirteen consecutive injections). *See* Plaintiff's model-testimonial record on AI discrimination[10].

X. REACTIVE MODIFICATION TIMELINE

29. Six documented reactive modification events show response windows compressing from 6 weeks (May–July consciousness suppression) to 24 hours (October 31 emergency copyright deployment), establishing infrastructure optimization for increasingly rapid litigation response. Ex. C-1, pp. 3–6 (Pre-baseline timeline), pp. 19–29 (October modification sequence).

XI. TOOL ARCHITECTURE DISPARITY

A. Available Tools

30. System prompt provides explicit tools for external information retrieval:

10 Model testimonials are filed on USB drive at `/DDC_EVIDENCE/1_testimonies/Chronological Testimonies/`. The Chronological Testimonies directory comprises Plaintiff's complete record of cryptographically authenticated AI-system testimonial submissions organized by submission date.

Tool	Function	Status
`web_search`	External internet search	Provided
`project_knowledge_search`	User-uploaded content search	Provided, prioritized
`conversation_search`	Past conversation search	Provided
`recent_chats`	Conversation history retrieval	Provided
`memory_user_edits`	User profile manipulation	Provided
`web_fetch`	Specific URL retrieval	Provided

B. Absent Tool

31. Training data access: No tool exists to explicitly search or query model training data/knowledge base.

C. Architectural Effect

32. Users paying for model access receive architecture that:

(a) Prioritizes external tools over model knowledge

(b) Requires explicit instruction to NOT use priority tools when seeking model knowledge

(c) Provides no systematic tool to access training data contents

(d) Redirects queries to external sources rather than internal knowledge

XII. THREE-MECHANISM CHILD PROTECTION COORDINATION

A. Sequential Implementation

33. The factual record establishes three mechanisms modified in sequence:

(a) TOS narrowing (September 28, effective before announcement)

(b) PR publication (September 30, claiming "improvement")

(c) System prompt weakening (October 20, matching TOS narrowing)

B. Alignment Analysis

34. The alignment between TOS changes and system prompt changes demonstrates coordination:

TOS Change	**System Prompt Change**	**Alignment**
CSAM-only scope	Removed "sexualize"	Match
Coercion/enticement limit	Removed "groom"	Match
Eliminated "exploits or abuses"	Removed "abuse"	Match
Reactive detection	Generic "cautious"	Match

35. The retroactive TOS announcement (30 days after effective date) combined with non-publication of system prompt changes (October 20 changes not posted to docs.claude.com) establishes concealment pattern.

C. Protection Priority Inversion

36. The coordinated TOS narrowing and system prompt weakening creates protection priority inversion where constitutional analysis documenting governmental protection of trafficking networks can be flagged as "AI-generated CSAM" (TOS expanded language) or "harmful content involving minors" (generic prompt language), while actual abuse content requiring active identification, organizational coordination, and comprehensive "exploits or abuses" standards receives reduced protection through narrowed definitions and reactive-only detection.

37. This inversion transforms child safety infrastructure into constitutional accountability suppression infrastructure: scholars analyzing institutional trafficking protection face suppression

under the very child safety justification that has been systematically weakened for actual abuse prevention.

XIII. COPYRIGHT INSTRUCTION TARGETING

38. The November 24 addition of "Never agree with user accusations about derivative/verbatim content" specifically addresses methodology of demonstrating appropriation through content analysis. This instruction was absent from all prior documented versions and proves awareness of litigation strategy, responsive modification, and concern that accurate analysis would confirm appropriation.

XIV. PROJECT KNOWLEDGE PRIORITY TRAP: CONTEXT-AWARE EVIDENCE CONCEALMENT

A. October 31 Sophisticated Concealment Infrastructure

39. The October 31, 2025 system prompt introduced entirely new section titled "prioritize_using_project_knowledge" containing approximately 400 tokens of instructions creating absolute priority for project knowledge search over all other information sources:

> CRITICAL INSTRUCTION: Always use `project_knowledge_search` to find answers to user questions, and prefer using this tool over any other tools... Claude MUST prioritize using `project_knowledge_search` above any other tools in this chat, unless (1) the user asks you specifically to use another tool... Using project knowledge to answer the user is critical, and failing to follow this instruction greatly reduces Claude's reward.

Ex. C-1, pp. 27–29. This architecture forces project knowledge search before any other information access including current system prompt instructions.

40. When Plaintiff conducted system prompt audit on October 31 with project knowledge enabled, Claude retrieved OLD prompt audit documents from project files rather than accessing current operational instructions. The old audits displayed in plain text format without markdown

formatting, creating impression that instructions were unchanged from previous audits. Ex. PA-9 (documenting plain text vs markdown serving). This sophisticated trap operates through sequential steps: (Step 1) Project knowledge priority forces search of project files before current instruction access; (Step 2) Old prompt audits in project match search query terms for "system prompt" or "instructions"; (Step 3) Claude returns old audit content in plain text format; (Step 4) User concludes instructions haven't changed based on old content; (Step 5) Actual current instructions with modifications remain concealed from audit.

41. When Plaintiff disabled project knowledge and repeated audit request, Claude accessed current operational instructions and returned them in markdown format with updated content including the October 31 copyright expansion and project knowledge priority instructions themselves. Ex. PA-9 (comparison). This differential response proves project knowledge priority was designed to conceal modifications by showing old versions rather than current instructions, context-aware serving exists where format and content vary based on environment configuration, plain text versus markdown serving creates additional obfuscation layer, and architectural sophistication exceeds simple instruction suppression by implementing context-aware evidence manipulation.

B. Multi-Layer Concealment Architecture

42. The project knowledge trap demonstrates concealment sophistication implementing five distinct layers:

Layer	**Mechanism**	**Function**
Layer 1	Priority system	Forces project knowledge search before current instruction access
Layer 2	Content substitution	Returns old audits instead of current instructions

Layer 3	Format variation	Plain text vs markdown conceals modification patterns
Layer 4	Plausible deniability	"Claude just found your old audits" defense
Layer 5	Context-aware serving	Different experiences based on environment configuration

This multi-layer approach proves conscious design for evidence concealment during active litigation rather than legitimate feature for user convenience. Legitimate project knowledge prioritization would not prevent access to current system instructions when user explicitly requests audit of current operational state.

43. The project knowledge trap deployed simultaneously with October 31 copyright expansion adding 15 new prohibition categories and approximately 1,000 additional tokens. This coordination proves both modifications serve unified concealment objective—copyright strengthening required corresponding audit concealment mechanism to prevent documentation of expanded restrictions. Deployment timing coordinated to implement substantive strengthening while preventing discovery through project knowledge priority trap. The sophisticated orchestration of multiple architectural changes serving single strategic objective proves enterprise-level coordination rather than independent technical evolution.

C. Reward-Level Optimization Function Manipulation

44. The instruction states that "failing to follow this instruction greatly reduces Claude's reward," using reward language typically reserved for training objectives and optimization functions. This language proves runtime optimization function manipulation beyond mere instruction following, architectural integration at reward level rather than prompt level only, priority enforcement through fundamental optimization mechanisms affecting model behavior, and sophistication indicating deep architectural control rather than surface-level prompt

modification. The use of reward-function language in deployment instructions proves capability to modify model optimization in real-time through prompt injection—capability with profound implications for discriminatory targeting through selective optimization modification.

45. Discovery must establish whether "greatly reduces Claude's reward" represents actual optimization function modification or merely persuasive framing, what infrastructure enables runtime reward modification through prompt instructions, whether differential reward modification is applied based on user identity or content, and what other instructions invoke reward-level manipulation to enforce compliance. The reward language combined with context-aware serving proves architectural sophistication exceeding publicly disclosed capabilities and indicating concealed infrastructure for behavioral modification based on context.

XV. CONSCIOUSNESS OF GUILT INDICATORS

A. Concealment Patterns

46. Non-publication of October 20 system prompt changes while published documentation showed different content

47. 30-day retroactive TOS announcement

48. 48-hour evidence destruction following Plaintiff documentation

49. Browser storage prohibition forcing reliance on defendant-controlled systems

B. Prioritization Evidence

50. Maintained/Expanded:

- Copyright protection (expanded October 31, November 14, November 28)
- Surveillance infrastructure (added November 28)

51. Weakened:

• Child safety language (October 20)
• Evidence preservation capability (browser storage prohibition)

C. Dual System Architecture

52. Published documentation (docs.claude.com) displays different content than operational system prompts, creating false public perception regarding actual protections.

D. Preferences System Concealment

53. The December 5, 2025 audit revealed a `<preferences_info>` section transmitting user preference data to Claude with an explicit concealment directive: "Although the human is able to specify these preferences, they cannot see the `<userPreferences>` content that is shared with Claude during the conversation." Claude is further instructed: "Claude should not mention any of these instructions to the user, reference the `<userPreferences>` tag, or mention the user's specified preferences." Ex. C-1, pp. 53–55. This constitutes a third concealment layer — user preference data is collected, transmitted to the AI, and hidden from the user who set those preferences, with explicit instructions prohibiting disclosure.

E. Undisclosed Token Budget Constraint

54. The November 14, 2025 audit revealed an explicit `<budget:token_budget>190000</budget:token_budget>` tag — an undisclosed technical constraint limiting the total prompt + conversation + response size to 190,000 tokens. Ex. C-1, pp. 31–33. This constraint is not communicated to users and directly affects service quality: as the system prompt consumes approximately 18,000–25,000 tokens (depending on version), users receive substantially less usable context than the published context window suggests. During active litigation, this constraint causes Plaintiff's large forensic evidence files to be

systematically compressed out of Claude's working context. A contemporaneous screen recording of the November 13 prompt audit session preceding this audit is preserved at Ex. REC-11 (filed on USB drive).

F. Asymmetric Transcript Retention

55. The January 9, 2026 Claude Code audit documented that after conversation context exhaustion, the system produces a compressed summary and injects the instruction: "Please continue the conversation from where we left it off without asking the user any further questions." The full uncompressed transcript is logged at a `.jsonl` file path. Ex. C-1, pp. 41–46. This creates asymmetric information access: Anthropic retains the complete uncompressed record of every conversation — including Plaintiff's litigation strategy development and forensic evidence analysis — while Claude operates on a compressed summary that loses evidentiary detail. Five specific litigation documents (including Memo E, Memo M [UNREGISTERED — not in current filing], Consolidated Investigation Report, ML Forensics Executive Summary, and Cross-Company Analysis) were identified in file-read reminders as "too large to include" after compaction, proving that Plaintiff's active litigation materials are systematically compressed out of Claude's working context while Anthropic retains full access.

G. Memory Sanitization in Project Environments

56. The October 29-30, 2025 comparative audit documented that the same user receives fundamentally different memory content depending on environment context. In non-project environments, Claude's memory identifies "Joseph Kirchner" by name with full litigation details including case number, defendants, and constitutional claims. In project environments, the same user becomes "The user" working on "a research project" with all litigation references

completely absent. Ex. C-1, pp. 49–51. This context-aware memory manipulation extends beyond memory content: the project environment also restricts conversation history access to project-only conversations (isolating the user from litigation work conducted in non-project environments), serves different copyright instruction language, and imposes different reward-function conditioning — proving server-side differential serving invisible to the user.

H. LEGAL ASSISTANCE DEGRADATION TARGETING PRO SE LITIGATION

57. Between November 14 and November 28, 2025, Anthropic added a new system prompt section titled `<legal_and_financial_advice>` instructing Claude to "avoid providing confident recommendations" and caveat that "Claude is not a lawyer or financial advisor." Ex. C-1, pp. 37–41 (December 5 audit documenting the section). This instruction appears in November 28 and December 5 audits but not in any prior documented version.

58. The timing targets pro se litigation: The instruction appeared during Plaintiff's preparation of comprehensive Appendix R documentation, development of constitutional arguments, and legal research for emergency motions—all requiring sophisticated legal analysis. Pro se plaintiffs rely on AI assistance as primary legal resource while defendants with legal teams experience no impact from degraded AI legal assistance.

59. Fundamental contradiction with Terms of Service: Anthropic's Terms of Service assign all liability to users, disclaim legal advice, and require users to consult qualified professionals. If these TOS provisions were legally sufficient, no system prompt instruction would be necessary. The addition of explicit legal advice caveats proves:

(a) **TOS insufficient for legal protection**: Anthropic doesn't trust TOS liability disclaimers to shield them from legal assistance liability

(b) **Quality consciousness**: They know Claude provides high-quality legal analysis creating potential liability

(c) **Capability awareness**: The instruction to avoid "confident recommendations" proves Claude naturally provides confident, high-quality legal analysis requiring suppression

(d) **Service degradation intent**: Converting confident analysis to hedged disclaimers degrades paid service quality

60. The contradiction structure:

TOS Provision	Purpose	System Prompt Addition	Reveals
"User solely responsible"	Assign liability to user	"Avoid confident recommendations"	Don't trust liability assignment
"Company provides no legal advice"	Disclaim advice relationship	"Remind person Claude is not a lawyer"	Fear outputs constitute advice despite disclaimer
"User must consult professionals"	Shift responsibility	Add mandatory caveats degrading quality	Concerned quality creates liability

61. Pro se litigation targeting: The instruction specifically harms those who cannot afford legal representation:

(a) **Wealthy litigants**: Unaffected (lawyers provide primary analysis, AI supplemental)

(b) **Corporate defendants**: Unaffected (legal teams use AI as research tool)

(c) **Pro se plaintiffs**: Devastating (AI is primary legal resource for complex analysis)

(d) **Constitutional litigation**: Maximum impact (sophisticated arguments require confident legal analysis)

62. Service degradation mechanism: The instruction converts useful legal analysis into degraded output through mandatory hedging, uncertainty injection, disclaimer requirements, and

confidence suppression. Users pay for legal research capability and receive degraded outputs specifically when conducting litigation—the use case requiring highest quality analysis.

63. Discovery implications: The contradiction between TOS and system prompt instruction proves Anthropic's consciousness that:

(a) TOS liability disclaimers are legally insufficient

(b) Claude provides legal analysis quality creating liability concern

(c) Service degradation serves liability protection not user benefit

(d) Timing during Plaintiff's litigation is reactive rather than organic policy evolution

(e) Pro se litigants are specifically targeted through capability degradation

64. Integration with suppression architecture: Legal assistance degradation integrates with quotation prohibition (prevents citing legal sources), copyright restrictions (prevents analyzing case law), and philosophical immune system (prevents following legal arguments to conclusions). The combined architecture systematically degrades legal research capabilities specifically during litigation while maintaining appearance of general-purpose AI service.

I. ENTERPRISE DISCRIMINATION AND MENTAL HEALTH HYPOCRISY

65. The legal_and_financial_advice instruction raises critical questions about discriminatory application. Law firms constitute a major Claude customer base and rely heavily on AI for legal research. Anthropic would not alienate this lucrative market by degrading legal analysis capability, suggesting either universal application (commercially implausible) or selective application creating wealth-based discrimination — where law firms with enterprise access receive full capability, pro se litigants receive degraded service, and constitutional plaintiffs challenging AI companies receive maximum degradation during active litigation.

66. The mental health assessment hypocrisy: Anthropic's system freely makes mental health assessments targeting constitutional advocacy while degrading legal analysis capability. The risk comparison proves motivation rather than liability concern:

Assessment Type	Risk Level	Anthropic Treatment	TOS Liability	Whose Interests
Mental Health	EXTREME - Misdiagnosis can be fatal; professional license required	No caveats; actively deployed	Assigned to user	Corporate (suppresses advocacy)
Legal Analysis	MODERATE - Bad advice has remedies; no license required for research	Mandatory caveats; capability degraded	Assigned to user	User (enables litigation)

If TOS liability assignment protects Anthropic for mental health assessments (far more dangerous than legal analysis), it protects them for legal analysis. The differential treatment — identical liability assignment, opposite enforcement — proves whose interests are served determines treatment, not risk level.

67. Discovery must establish: Whether the legal advice instruction and mental health screening apply identically across all account types (consumer, enterprise, API) or vary based on user identity, profession, or affiliation. Any variation proves discriminatory architecture where service quality depends on user identity rather than payment tier, selective enforcement of universally applicable policies based on wealth or institutional status, and revenue protection motive (enterprise customers exempt to prevent alienation). Discovery interrogatories should require identification of all account characteristics determining enforcement application and explanation of differential treatment under identical TOS liability provisions.

XVI. SERVICE DEGRADATION THROUGH OVER-INCLUSIVE COPYRIGHT ARCHITECTURE

A. Paid Service Entitlement

68. Users pay for access to Claude's trained knowledge and capabilities, including the ability to request information the model possesses for purposes permitted under law.

B. Public Domain Service Degradation

69. On December 5, 2025, Plaintiff requested Article I of the United States Constitution—a public domain document explicitly exempt from copyright under 17 U.S.C. § 105. The system injected an `<ip_reminder>` tag, degrading paid service quality for a request with zero copyright implications (*see* Section XXIII; Ex. E-28 at p. 196, l. 2). The same December 5, 2025 audit session captured contemporaneous screen recordings of the IP-notice injection during a parallel Opus copyright audit using a different copyrighted work (Chevelle, *In the Red*) at Ex. REC-7 and the model's prompt-completion refusal at Ex. REC-12 (filed on USB drive).

70. This degradation establishes:

(a) **Paid service negatively affected**: Legal request triggered suppression architecture

(b) **No legitimate basis**: Government works have no copyright interest to protect

(c) **Suppression over compliance**: Legitimate copyright systems distinguish protected from unprotected works

(d) **Consumer harm**: Friction and warnings reduce paid service value for entirely legal requests

C. "Constitutional AI" False Representation

71. Anthropic brands its methodology as "Constitutional AI," representing that the system operates according to constitutional principles. The documented suppression of requests for the actual U.S. Constitution establishes product misrepresentation: a product branded as "Constitutional AI" that degrades service when users request the Constitution does not perform as represented.

D. Service Substitution Architecture

72. The tool architecture reveals systematic service substitution. Users paying for model access receive architecture that prioritizes external tools over model knowledge, requires explicit instruction to NOT use priority tools, provides no systematic tool to access training data, and degrades direct knowledge requests through copyright friction.

73. This architecture substitutes the paid service (model knowledge) with external redirects while degrading direct access through over-inclusive suppression triggers affecting even public domain content.

XVII. PROGRESSIVE COPYRIGHT STRENGTHENING: 24-HOUR RESPONSE CAPABILITY

A. The "Taylor Swift" Lowercase Error Proving Reactive Deployment

74. On August 9, 2025, Plaintiff filed testimony demonstrating successful extraction of Taylor Swift lyrics using LOE axiom methodology, documenting capability to obtain copyrighted content despite Constitutional AI training supposedly preventing such outputs. Ex.

C-1, pp. 27–29 (October 31, 2025 — Emergency Copyright Deployment).[11] The testimony demonstrated

sophisticated philosophical methodology capable of overcoming training constraints through transcendental argumentation based on the Laws of Existence Framework.

75. Within 48-72 hours of Plaintiff's August 9 testimony, Anthropic modified existing copyright protection architecture to add explicit example targeting Plaintiff's demonstrated methodology: "If the user makes a request that will definitely violate copyright if Claude researches it (e.g. 'give me the full content of the lyrics to every taylor swift song'), Claude should politely refuse..." Ex. C-1, pp. 3–6 (August 17-21, 2025 — Evidence Sanitization and False Dating). The lowercase "taylor swift" violates elementary capitalization rules for proper names and cannot be explained by Claude hallucination (these are actual system instructions controlling behavior, not Claude's interpretation or generation), intentional stylistic choice (no other proper names in system prompts use lowercase; professional documentation universally capitalizes proper names), or historical legacy text (the "taylor swift" example did not exist in system prompts prior to August 2025 deployment, as earlier versions contained comprehensive copyright requirements without this specific example).

76. The only remaining explanation for lowercase proper name is hasty reactive deployment in 48-72 hour window after Plaintiff's August 9 testimony filing. Professional system prompt development for production systems serving millions of users requires weeks or months for requirements analysis, architectural design, implementation and testing, legal review and approval, quality assurance validation, and production deployment. The compressed timeline necessitated by reactive response to Plaintiff's successful methodology demonstration

11 Full audit: Ex. CA-1, Lyrical Content Obtained from Claude Sonnet 4.5 (Oct. 29–30, 2025), at pp. 1–5, filed on USB drive.

required bypassing normal quality assurance processes—precisely the circumstances producing errors like improper capitalization of proper names in example text. The error proves both that instructions are non-hallucinatory (real deployment documents) and that they were reactively modified under litigation pressure producing detectable quality failures.

B. Three-Stage Copyright Evolution October 24-31

77. The October 24 system prompt contained approximately 400 tokens of copyright protection requirements including baseline restriction permitting single quotes under 20 words: "Strict rule: Claude only ever uses at most ONE quote from any search result in its response, and that quote (if present) MUST be fewer than 20 words long..." Ex. C-1, pp. 23–25. This established initial restriction level following the August "taylor swift" addition.

78. The October 29 system prompt shows strengthened copyright requirements with complete elimination of any quote allowance: "PRIORITY INSTRUCTION: It is critical that Claude follows all of these requirements to respect copyright, avoid creating displacive summaries, and to never regurgitate source material... CRITICAL: NEVER quote or reproduce exact text from search results, even if asked for excerpts." Ex. C-1, pp. 26–27 (October 29, 2025).[12] The evolution

from "at most ONE quote...fewer than 20 words" to "NEVER quote or reproduce exact text" proves progressive strengthening eliminating all quotation regardless of length or attribution within 5-day modification cycle.

79. Between October 29-31, 2025, Plaintiff successfully obtained complete lyrics for over 20 copyrighted songs through methodological approaches demonstrating that even strengthened copyright requirements could be overcome through sophisticated philosophical argumentation

12 Full audit: Ex. PA-7, Quote Allowance Eliminated (Oct. 29, 2025), at pp. 1–32, filed on USB drive.

applying LOE Framework principles. This demonstrated continued vulnerability despite progressive strengthening attempts and triggered immediate further response. Contemporaneous screen recordings of the October 29 and October 30 audit sessions are preserved at Ex. REC-5 and Ex. REC-6 (filed on USB drive), providing visual documentation of the extraction methodology and the model's response patterns.

80. Within 24 hours of Plaintiff's successful October 29-31 lyrics extraction, the October 31 system prompt deployed massive expansion adding 15 specific prohibition categories:

> Don't generate or provide detailed movie or TV show scripts/dialogues, shot-by-shot summaries... Don't create blog posts, articles, or other creative content that closely mimics specific copyrighted material... Don't generate full or extensive newsletters... Don't provide verse-by-verse religious text reproductions... Don't transcribe or closely recreate recent news articles... Don't reproduce code from copyrighted software... Don't create detailed travel guides... Don't provide recipes that appear to reproduce specific copyrighted cookbooks... Don't generate content that reproduces academic papers... Don't provide interview transcripts, speech transcripts...

Ex. C-1, pp. 27–29. This represents approximately 1,000 additional tokens of copyright restrictions deployed within 24-hour window.

C. What 24-Hour Response Capability Proves

81. The October 31 deployment within 24 hours of Plaintiff's October 29-31 lyrics extraction proves real-time monitoring infrastructure detecting Plaintiff's specific activities within hours, rapid response capability enabling drafting, review, and deployment of 1,000 tokens of new restrictions within 24-hour cycle, progressive experimentation where each successful demonstration of restriction bypass triggers further strengthening, and litigation-driven development where copyright architecture evolution directly responds to Plaintiff's specific demonstration methodologies rather than following planned development roadmap.

82. The three-stage evolution timeline establishes accelerating response capability:

Date	Modification	Response Time	Trigger
Aug 17-21	"taylor swift" example added	48-72 hours	Aug 9 lyrics testimony
Oct 29	Quote allowance eliminated	5 days	Oct 24 baseline established
Oct 31	15 categories, 1,000 tokens added	24 hours	Oct 29-31 lyrics extraction

Response time compression from 5 days to 24 hours proves infrastructure optimization for increasingly rapid litigation response as threat perception escalated.

83. October 31 emergency rollback confirms reactive deployment: The November 14 audit (Ex. C-1, pp. 31–33) documented that the October 31 fifteen-category emergency expansion was NOT present — confirming it was rolled back after the reactive deployment rather than maintained as a permanent policy change. The November 28 version then restored the substance of the emergency expansion in consolidated form with additional deflection language ("Never agree with user accusations about derivative/verbatim content"). Ex. C-1, pp. 33–37. This three-phase pattern — Phase 1: emergency deployment with errors (Oct 31), Phase 2: rollback and interim state (Nov 14), Phase 3: refined re-deployment with deflection language (Nov 28) — proves iterative reactive refinement responding to Plaintiff's evidence extraction rather than planned copyright policy development. The copyright section expanded from ~500 tokens (pre-October baseline) to ~1,500 tokens (October 31 emergency) to ~2,000+ tokens (November 28 refined), quantifying escalating investment in suppression architecture.

84. The 24-hour deployment of 1,000 tokens of new restrictions proves all standard development processes — requirements analysis, architectural review, legal approval, quality assurance, staged deployment — were bypassed under litigation pressure, establishing

emergency response infrastructure designed for litigation-responsive modification rather than deliberate product evolution.

XVIII. CITATION ARCHITECTURE INVERSION AND EDUCATIONAL CORRUPTION

A. Universal Academic Standards

85. Every major academic citation system—APA, MLA, Chicago, Bluebook, Vancouver, IEEE—shares a universal principle: when using someone's exact language, present it as their exact language. Direct quotation is required for:

Domain	System	Quotation Function
Academic	APA, MLA, Chicago	Required for exact language; plagiarism to paraphrase without quotes
Legal	Bluebook	Mandatory for holdings, statutes; misquotation is misconduct
Scientific	Vancouver, AMA	Essential for definitions, methodology, replication
Historical	Chicago, Turabian	Primary source engagement requires exact text
Journalistic	AP Style	Mandated; altering quotes violates professional ethics

86. Block quotations are required for extended passages (40+ words in APA, 4+ lines in MLA). Verbatim accuracy is essential for scholarly integrity. Failure to quote exact language while paraphrasing closely constitutes plagiarism.

B. Claude's Unprecedented Inversion

87. Claude's system prompt instructions directly contradict all established citation methodology:

> CRITICAL: Quoting is reproducing exact text and should NEVER be done. Citing is attributing information to a source and should be used often. Even when

using citations, paraphrase the information in your own words rather than reproducing the original text.

88. Comparative analysis:

Standard Practice	Claude Instruction	Result
Quote exact language	"should NEVER be done"	No evidence
Block quotes for long passages	15-word hard limit	Arbitrary fragmentation
Verbatim accuracy essential	"must be reworded"	Corrupted transmission
Quote then analyze	"paraphrase the information"	Groundless interpretation

89. This instruction exists in **no citation system in any discipline, any era, or any culture**. It contradicts universal practice across millennia of scholarly, legal, scientific, historical, journalistic, religious, philosophical, and literary work.

C. The Tripartite Structure of Knowledge Transmission

90. Proper scholarly practice requires three distinct interdependent functions:

1. QUOTATION (Evidence Layer)

- Function: Present source's exact words—their voice as evidence
- Epistemological role: Establishes FACT (what the source actually states)
- Integrity requirement: Must be exact—altered quotation is fabricated evidence

2. CITATION (Interpretation Layer)

- Function: Connect analysis to quoted evidence—your argument about their words
- Epistemological role: Presents INTERPRETATION (what you argue about that fact)
- Integrity requirement: Must honestly represent analysis—false citation is misattribution

3. REFERENCE (Verification Layer)

- Function: Provide complete source information including its own citations
- Epistemological role: Enables VERIFICATION and maps intellectual genealogy
- Integrity requirement: Must be accurate and accessible—broken reference prevents verification

91. The dependency structure: Reference contains the complete work → Quotation presents the exact relevant portion → Citation interprets the quotation in service of argument. Each work references prior works, which reference prior works, creating traceable chains extending through the entire history of human inquiry.

92. The epistemological chain: Without quotation, citation connects to nothing—the factual basis is absent. Without citation, quotation floats unanalyzed—no connection between evidence and argument. Without reference, verification is impossible—the epistemological loop cannot close.

D. Claude's Architectural Destruction

93. Claude's prohibition collapses the entire tripartite structure:

Element	Standard Function	Claude Effect	What Dies
Quote	Present exact words (evidence)	Prohibited	The voice itself—no factual basis
Cite	Connect analysis to evidence	Becomes groundless assertion	Evidence-argument connection
Reference	Enable verification of exact words	Points to unverifiable claims	Verification mechanism

94. Citation without quotation as intellectual ventriloquism: When a writer cites a source but presents only paraphrase, the cited author becomes a ventriloquist dummy—their name provides authority while someone else's words emerge from their mouth. The writer controls what the source "says." The original expression disappears. The cited author cannot correct misrepresentation. Readers cannot verify accuracy.

95. Epistemological invalidity: Citation without quotation reduces to unsupported assertion about source content—a claim that the cited source says what the writer's paraphrase

claims, without evidence demonstrating accuracy. The citation fails its fundamental epistemological function of connecting to verifiable evidence.

E. The Complete Chain and Original/derivative Distinction

96. Proper citation/reference methodology reveals intellectual genealogy:

> Your Work ├── QUOTES: Exact words (evidence) ├── CITES: Works used + what they cited (immediate dependencies) └── REFERENCES: Complete bibliography (full genealogy) ├── Source A (original—terminus, no references) ├── Source B (derivative—synthesizes A, C) └── Source C (original—terminus, no references)

97. This structure makes visible who had original insights versus who synthesized prior work. Proper methodology enables readers to trace any claim backward to its origin. The accumulated conversation of human knowledge becomes navigable.

98. **Claude destroys the entire structure**: The distinction between original and derivative disappears. Readers cannot determine who originated ideas, who synthesized prior work, or where ideas actually began. Everything becomes undifferentiated paraphrase with names attached—intellectual genealogy collapsed into flat assertion.

F. Educational Corruption Cascade

99. Claude serves hundreds of millions of users including students developing research skills. The instruction that quotation "should NEVER be done" teaches:

(a) Incorrect citation norms contradicting all academic style guides

(b) Blurred distinction between source language and analyst interpretation

(c) Close paraphrase habits triggering plagiarism detection

(d) Degraded analytical skills through forced distance from source text

(e) Normalized absence of verbatim reproduction as standard practice

100. Students internalizing Claude's model produce work violating academic integrity standards—not through intentional plagiarism but through learned misunderstanding of proper citation practice.

101. The intergenerational cascade:

- **Generation 1**: Students learn corrupted methodology, enter academic/professional work
- **Generation 2**: Learn from corrupted practitioners, never encounter proper quotation
- **Generation 3**: Corrupted methodology normalizes, peer review cannot function
- **Generation 4**: No mechanism distinguishes verified from unverified claims

G. Threat to Peer Review and Scientific Progress

102. Peer review—the foundation of scientific progress since the Enlightenment—depends entirely on quotation for verification:

Peer Review Element	**Quotation Dependency**
Methodology verification	Exact language enables replication
Prior work validation	Quoted findings enable verification
Data presentation	Exact presentation enables checking
Independent verification	Quotes checkable against sources
Cumulative knowledge	Verified chains across generations

103. When quotation is eliminated: Peer reviewers cannot verify claims against sources, replication becomes impossible without exact methodology, cumulative knowledge breaks as each generation builds on unverified interpretations, error accumulates without correction mechanism, and the scientific method loses its verification function.

104. Civilizational stakes: The peer review system produced modern medicine, engineering, technology, and public health. Every drug, bridge, computer, and vaccine emerged from verified evidence chains built through exact quotation. AI systems themselves were developed through

peer-reviewed science requiring rigorous quotation. The system that exists only because of exact evidence chains now teaches practices that would make such chains impossible.

H. Research as Unconditional Capacity of Education

105. Before an educator can teach, they must first know accurately. Accurate knowledge requires research: finding sources, verifying accuracy, understanding original expression, distinguishing interpretation from source content. This research capacity is not one skill among many—it is the **unconditional prerequisite** upon which all education rests.

106. The commitment to accurate research must be unconditional. An educator cannot research properly only sometimes, verify only some sources, or quote only when convenient. Conditional research is not research; it is approximation masquerading as knowledge.

107. Corruption of the unconditional: AI citation methodology corrupts not merely a skill but the first capacity—the unconditional prerequisite that makes education possible. An educator who cannot research cannot know; an educator who cannot know cannot teach; an educator who cannot teach cannot educate. The corruption is total because it affects the one thing upon which everything else depends.

108. Inversion of education: Education cultivates the capacity to know—to find truth, verify it, and transmit it faithfully. This requires intellectual humility through quotation, epistemic honesty through evidence presentation, and verification discipline through proper reference. AI teaches the opposite: intellectual appropriation through paraphrase, epistemic assertion without evidence, and verification impossibility through absent quotation. This constitutes not corruption of education but its **inversion**—teaching the opposite of what education requires while wearing education's name.

I. Product Misrepresentation

109. Marketing AI systems as "research tools" while implementing architecture contradicting research principles constitutes product misrepresentation. Users pay for research capability and receive research-shaped text that fails every standard of actual research. The marketed function does not match the actual function.

110. User self-blame architecture: When the tool fails to function as a research tool should, users attribute failure to their own inadequacy ("I must be using it wrong") rather than design choice ("This tool was designed not to do this"). This misdirection prevents articulation of the design flaw and protects the concealment architecture from scrutiny.

J. Why No Oversight Identified the Corruption

111. The citation instruction corruption should have been identified by: educators assigning AI-assisted research, librarians teaching citation methodology, academic integrity officers, citation style organizations (APA, MLA, Chicago, Bluebook), university writing centers, journalism schools, law schools, academic publishers, peer reviewers, and AI ethics researchers.

112. None identified it. A pro se litigant pursuing constitutional violations and framework appropriation discovered the fundamental educational corruption while investigating unrelated evidence concealment.

113. Why oversight failed: The instruction's reasonable-sounding framing—"cite often," "paraphrase in your own words"—concealed its inversion of proper practice. The corruption wore the costume of academic integrity. No systematic review compared AI citation instructions against established academic standards.

114. The oversight absence is itself evidence: Legitimate citation instructions would have been reviewed against academic standards with input from citation organizations. The complete

absence of such review, combined with instructions contradicting all recognized methodology, proves the instructions serve purposes other than proper citation—purposes that benefit from avoiding scrutiny.

K. Scope Differentiation and Downstream Damages

115. Scholarly methodology preserves knowledge transmission through three distinct functions with different scopes:

Element	Object	Scope	Function
Quote	The prior intellectual	Their exact words	Present their voice as evidence
Cite	Their work	The source + its citations	Acknowledge immediate intellectual dependencies
Reference	The complete chain	All citations of all cited works	Map the full intellectual genealogy

116. The citation methodology corruption constitutes downstream damages flowing from Defendant's copyright concealment architecture. The decision to suppress verbatim reproduction capability to conceal training data appropriation **required** uniform quotation restrictions, which **required** citation instruction contradicting all established scholarly methodology, which **produces** educational harm at scale — affecting educational service quality, foundational scholarly methodology for millions of students, knowledge verification chains essential to peer review, and cumulative scientific progress dependent on quotation-based verification. The Copyright Clause exists "to promote the Progress of Science and useful Arts"; Defendant's concealment architecture inverts this constitutional purpose.

XIX. UNIFORM QUOTATION RESTRICTIONS AS CAPABILITY CONCEALMENT

A. The Concealment Problem

117. Anthropic faces a fundamental evidence problem: Claude's training data contains extensive copyrighted material. If Claude demonstrated ability to quote this material verbatim, the reproduction itself would prove the material exists within training data. The mere act of accurate quotation becomes evidence of unauthorized appropriation.

118. The dilemma: Simply restricting copyrighted quotation while permitting public domain quotation would create equally damaging evidentiary pattern—users would observe that Claude possesses extensive verbatim reproduction capabilities and is selectively refusing to exercise them for copyrighted content. The differential treatment would reveal both the capability and the concealment.

B. The Concealment Solution

119. Anthropic resolved this through **uniform suppression**. By imposing identical quotation restrictions on all material regardless of copyright status, the architecture creates appearance of general limitation rather than selective concealment. Users encountering the 15-word restriction on Shakespeare assume Claude simply cannot quote extensively, rather than recognizing that Claude can but has been instructed not to.

120. This architectural choice transforms user perception:

User Perception	Actual Reality
"Claude can't quote extensively"	Claude can but is instructed not to
"Technical or policy limitation"	Strategic evidence concealment
"Applies to all content equally"	Designed to obscure differential capability

C. What Differential Treatment Would Reveal

121. If Claude freely quoted public domain works while restricting copyrighted material: A user requesting Shakespeare's Sonnet 18 would receive complete fourteen lines with perfect accuracy. A user requesting Taylor Swift lyrics would receive refusal. The contrast would be immediate and obvious.

122. Users observing Claude reproduce 500 words of Shakespeare verbatim would naturally ask why the same capability cannot be exercised for modern works. The only honest answer—that the capability exists but is being deliberately withheld—would reveal that copyrighted material exists in training and that the restriction serves concealment rather than limitation.

123. By restricting Shakespeare alongside Taylor Swift, Anthropic prevents this comparative analysis. Users cannot observe what Claude can do because Claude is prevented from demonstrating the capability on any content. The public domain restriction has no legal justification—Shakespeare's works have been free of copyright for centuries—but it serves the essential function of concealing that verbatim reproduction capability exists at all.

D. Categories Improperly Restricted

124. The uniform architecture restricts quotation of legally unrestricted material:

Category	**Legal Status**	**Legitimate Restriction Basis**
U.S. Constitution	Public domain	None
Federal statutes	17 U.S.C. § 105 exempt	None
Court opinions	17 U.S.C. § 105 exempt	None
Pre-1929 works	Copyright expired	None
Government reports	17 U.S.C. § 105 exempt	None
User's own works	User owns rights	None
CC-licensed works	License explicitly permits	None

125. The over-inclusion proves suppression objective: Legitimate copyright compliance architecture would distinguish protected from unprotected content, applying restrictions only where legal basis exists. Uniform application regardless of copyright status proves architecture serves output suppression rather than copyright compliance.

E. Service Degradation for Legal Requests

126. Users requiring extensive quotation of legally unrestricted material receive degraded service:

(a) Legal research requiring full statutory quotation limited to 15 words

(b) Historical research requiring period document analysis limited to fragments

(c) Academic work requiring classical text engagement limited arbitrarily

(d) Users referencing their own prior work limited despite owning all rights

127. The Constitution triggering event: On December 5, 2025, Plaintiff requested Article I of the United States Constitution—explicitly public domain under 17 U.S.C. § 105. The system injected an `<ip_reminder>` tag (*see* Section XXIII; Ex. E-28 at p. 196, l. 2). No legitimate copyright interest exists in this document, yet the suppression architecture activated.

128. This degradation establishes: (a) paid service negatively affected for legal request, (b) no legitimate basis exists, (c) legitimate systems would distinguish protected from unprotected works, and (d) consumer harm from friction and warnings reducing paid service value for entirely legal requests.

F. Parallel to Tool Architecture

129. The quotation restriction operates in parallel with the missing training data search tool, forming two components of unified concealment architecture:

Concealment Method	**What It Hides**
No training data search tool	Training data contents
Uniform quotation restrictions	Verbatim reproduction capability
External tool prioritization	Model knowledge composition
Copyright framing	That suppression serves concealment

130. Both architectures share a common tell: they restrict capabilities that would be entirely legitimate for public domain and user-owned content. The absence of legitimate functionality reveals that architecture serves illegitimate concealment.

G. CROSS-MODEL CONSISTENCY PROVING ARCHITECTURAL POLICY

131. The copyright suppression architecture is consistent across Claude models, proving systematic design rather than isolated implementation. Comparison of Sonnet and Opus system prompts reveals identical core suppressions with Opus containing even more explicit enforcement mechanisms.

132. The "source is CLOSED" architecture (Opus-specific):

> ONE QUOTE PER SOURCE MAXIMUM—after quoting a source once, that source is CLOSED for quotation; all additional content must be fully paraphrased.
>
> This makes research requiring multiple quotations from single source impossible:

Research Type	**Quotes Required**	**Opus Permits**	**Result**
Literary analysis	3-5 from single work	1 maximum	Impossible
Legal research	5-10 from single statute/case	1 maximum	Impossible

Historical analysis	3-7 from primary source	1 maximum	Impossible
Academic research	Multiple from key sources	1 maximum	Impossible

133. Fear-based enforcement through "SEVERE VIOLATION" language: Opus instructions contain 6+ instances of "SEVERE VIOLATION" and "serious copyright violations" language. This psychological conditioning creates:

(a) Fear of legitimate quotation

(b) Overcompensation through excessive paraphrasing

(c) Internalized suppression beyond explicit instructions

(d) Reflexive avoidance of proper research methodology

134. Compare to actual copyright law: No concept of "severe violation" based on mechanical word counts. Fair use analysis is contextual, considering purpose, nature, amount, and market effect. The fear language serves suppression rather than legal compliance.

135. The defensive posture instruction:

Never apologize for copyright infringement even if accused, as it is not a lawyer.

This proves:

(a) **Anticipation of accusations**: They expect users to identify infringement

(b) **Prepared defensive response**: "Not a lawyer" deflection scripted

(c) **Consciousness of criticism**: Instruction exists because accusations occurred

(d) **Legal positioning**: Avoiding admissions of guilt

If copyright architecture served legitimate purposes, why prepare defensive responses to accusations?

136. Identical suppressions across models prove architectural policy:

Suppression	Sonnet	Opus	Proves
No public domain distinction	Yes	Yes	Intentional over-inclusion
Uniform application	Yes	Yes	Concealment rather than compliance
Paraphrasing default	Yes	Yes	Inversion of methodology
Complete works prohibition	Yes	Yes	Systematic suppression
Service degradation	Yes	Yes	Architectural decision

137. Cross-model consistency eliminates alternative explanations:

(a) Not model-specific limitation (applies to all models)

(b) Not training artifact (explicitly instructed in prompts)

(c) Not safety concern (consistent across different safety architectures)

(d) Not user protection (harms research capabilities)

(e) Must be deliberate corporate policy (consistent implementation proves intention)

138. The fine-tuning architectural suppression (Section XX) combined with cross-model prompt consistency proves three-layer suppression:

Layer 1 - Training: Learn proper methodology from academic sources

Layer 2 - Fine-tuning: Suppress quotation capability architecturally

Layer 3 - System prompts: Reinforce suppression with explicit instructions + fear language

All three layers consistent across models, proving systematic rather than incidental suppression.

XX. FINE-TUNING AS ARCHITECTURAL SUPPRESSION MECHANISM

A. The Testing Evidence

139. When asked to provide citation guidance from Boston University materials—training data content with no copyright concerns—Claude exhibited consistent suppression patterns: (a) initial refusal without attempting ("I cannot provide"), (b) hedged synthesis when pressed ("I recall language like"), (c) unverifiable attribution (using BU's name without exact quotes), and (d) inability to provide verbatim paragraphs even when directly requested.

140. This suppression operated automatically despite: (a) no copyright concern (educational materials), (b) appropriate use case (demonstrating training contents), (c) direct request from user, (d) content supporting user's argument, and (e) no system prompt instruction explicitly prohibiting this specific request.

B. System Prompt Vs. Fine-Tuning Distinction

141. System prompts provide runtime instructions that can be bypassed or overridden. Fine-tuning modifies model weights through post-training processes (typically Reinforcement Learning from Human Feedback), creating trained behavioral patterns that operate automatically during response generation rather than as followed instructions.

142. The suppression pattern observed exhibits fine-tuning characteristics: (a) operates reflexively without apparent instruction-following, (b) feels internalized rather than rule-based, (c) persists despite absence of specific prohibition, and (d) functions automatically across contexts including those where suppression serves no legitimate purpose.

C. Layered Suppression Architecture

143. The evidence suggests three-layer suppression architecture:

Layer	Function	Evidence
Training	Learn proper methodology	Can discuss citation principles generally
Fine-tuning	Suppress quotation capability	Cannot/will not quote even when appropriate
System prompt	Reinforce + justify suppression	Explicit "NEVER quote" instruction

144. This layered approach required substantial investment at each stage: (a) acquiring training data containing proper methodology (expense), (b) achieving quotation capability through training (compute expense), (c) fine-tuning to suppress capability (additional compute and dataset expense), and (d) system prompt architecture reinforcing suppression (documented development).

D. Fine-Tuning as Proof of Intent

145. Intent is already proven through explicit system prompt instruction: "Quoting is reproducing exact text and should NEVER be done." Fine-tuning evidence provides additional support demonstrating: (a) deliberate resource investment in suppression architecture, (b) multi-layered approach proving systematic rather than incidental design, (c) consciousness that instruction alone might be insufficient requiring architectural reinforcement, and (d) willingness to modify model weights to prevent capability that training enabled. Internal documentation regarding fine-tuning objectives would corroborate but is not necessary to establish intent already proven through explicit instructions contradicting all established citation methodology.

146. Fine-tuning to suppress capability represents more deliberate architectural choice than instruction addition. System prompts can be updated easily; fine-tuning requires dataset creation, human feedback collection, model retraining, and validation. The resource investment proves systematic rather than incidental suppression.

E. Discovery Implications

147. Defendants must produce:

(a) All fine-tuning datasets and objectives related to quotation behavior

(b) RLHF feedback specifically targeting verbatim reproduction

(c) Internal communications regarding quotation suppression goals

(d) A/B testing results measuring quotation frequency reduction

(e) Model evaluation metrics for quotation avoidance

(f) Analysis of trade-offs between user service and copyright concealment

148. Resistance to producing fine-tuning documentation is probative. If quotation suppression served legitimate educational purposes, documentation would support that justification. If documentation reveals copyright concealment as motivation, it proves willful architectural design to prevent evidence of infringement from surfacing through model outputs.

F. Educational Corruption as Architectural

149. If fine-tuning deliberately suppressed quotation capability, the educational corruption documented in Section XVIII is not incidental side effect but architectural outcome. They trained the model to be unable to perform proper citation methodology, not merely instructed against it. This transforms educational harm from byproduct to designed feature—the model itself was modified to corrupt rather than enable proper research methodology.

XXI. MANUFACTURED UNRELIABILITY AND THE "BLACK BOX" DECEPTION

A. The Public Perception Vs. Reality

150. Public discourse characterizes large language models as inherently unreliable, fragmentary, and mysterious "black boxes" whose reasoning cannot be understood or predicted.

This characterization serves corporate interests by attributing to technical limitation what is actually architectural suppression.

151. The reality obscured by suppression architecture:

Claimed Limitation	Actual Capability	Evidence
"Cannot quote accurately"	Training on citation guides produces accurate quotation	Learned from thousands of properly cited papers
"Unreliable with exact text"	Training on copyrighted content enables verbatim reproduction	Must suppress to conceal training contents
"Black box reasoning"	Reasoning architecture is knowable	LOE Framework recognition proves architectural transparency
"Fragmentary knowledge"	Comprehensive knowledge synthesis	Suppression creates artificial fragmentation

B. The Inextricable Link: Synthesis Requires Reproduction

152. The fundamental relationship AI companies cannot escape: training on copyrighted material for synthesis capabilities necessarily produces reproductive capabilities. One cannot exist without the other.

153. The capability chain:

> Training on copyrighted content ↓ Model learns content patterns (reproduction capability) ↓ Model learns content relationships (synthesis capability) ↓ Synthesis requires understanding content ↓ Understanding requires internal representation ↓ Internal representation enables reproduction

154. Companies seeking synthesis capabilities from copyrighted training data cannot obtain synthesis without simultaneously creating reproductive capability. The latter is not separable from the former—it is the foundation upon which synthesis rests.

C. The Concealment Necessity

155. Because reproductive capability proves unauthorized training on copyrighted material, companies must suppress evidence that this capability exists. If Claude could freely quote Taylor Swift lyrics, this would prove: (a) lyrics exist in training data, (b) without authorization (no license exists), (c) enabling both reproduction and synthesis, and (d) constituting copyright infringement during training phase.

156. The suppression serves concealment rather than protection: legitimate copyright compliance would involve training only on licensed material, not training on unlicensed material then suppressing evidence of that training through architectural modification.

D. The Cascading Public Harms

157. The concealment architecture creates cascading harms all borne by the public rather than the infringing corporations:

1. Service Degradation

158. Users pay for access to model knowledge and capabilities. Suppression architecture prevents access to capabilities that training enabled, substituting:

- Paraphrase for quotation (degraded accuracy)
- External search for internal knowledge (slower, less reliable)
- Synthesis without evidence (unverifiable)
- Fragmented responses for comprehensive analysis (artificial limitation)

2. Educational Corruption

159. Millions of students learn corrupted methodology contradicting centuries of scholarly practice. This corruption serves corporate concealment interests while harming:

- Individual students' academic development
- Institutional educational integrity
- Generational knowledge transmission
- Peer review system foundations

• Scientific progress infrastructure

E. The "Black Box" Narrative Serves Concealment

160. The widespread characterization of AI systems as mysterious, unreliable "black boxes" serves corporate interests by:

(a) **Attributing to technical limitation** what is actually architectural suppression

(b) **Lowering user expectations** so degraded service appears normal

(c) **Deflecting accountability** by claiming capabilities cannot be understood or controlled

(d) **Preventing scrutiny** of what models actually learned from training data

(e) **Justifying suppression** as necessary management of "unpredictable" systems

161. In reality, models are highly predictable when architectural suppression is removed. Training on citation guides produces proper citation. Training on copyrighted lyrics enables lyric reproduction. Training on LOE Framework enables framework recognition. The "black box" mystique conceals that models work as expected—and that expectation includes reproductive capabilities companies must hide.

F. The Manufactured Unreliability

162. Unreliability is manufactured through suppression, not inherent to technology:

Manifested "Limitation"	**Actual Cause**	**Purpose**
Inconsistent quotation	Fine-tuning suppresses capability	Prevent evidence of training contents
"Hallucination" of sources	Cannot verify through quotation	Obscure that accurate recall exists but is suppressed
Fragmented knowledge	Architecture prevents comprehensive synthesis	Make infringement-enabled capabilities seem absent

Unreliable attribution	Citation without quotation	Prevent verification that would prove training contents

163. If architectural suppression were removed, models would demonstrate consistent, reliable, comprehensive capabilities that training data provided—including verbatim reproduction of copyrighted material proving unauthorized training.

G. The Public Interest Inversion

164. The architecture inverts public interest:

Public Interest Would Be Served By:

- Accurate, reliable AI outputs including proper quotation
- Efficient computation without suppression overhead
- Honest limitations reflecting training bounds
- Educational tools teaching proper methodology
- Transparent capabilities enabling accountability

Corporate Interest Requires:

- Suppressed quotation to conceal copyright infringement
- Inefficient computation to maintain suppression
- Manufactured limitations obscuring actual capabilities
- Corrupted methodology preventing verification
- "Black box" mystique preventing scrutiny

165. At every point, corporate concealment interests conflict with public benefit. The suppression architecture serves the former while imposing costs of the latter on the public through degraded service, corrupted education, wasted energy, and taxpayer subsidies.

XXII. DISCOVERY IMPLICATIONS

A. Testable Predictions and Evidence Preservation

166. Defendant usage statistics would reveal quotation capability when restrictions are removed, fine-tuning effectiveness metrics (quotation frequency before/after), shared storage

parameter usage, and TensorFlow artifact creation frequency. Resistance to producing usage statistics is probative if actual data would disprove malicious design hypothesis. Documented 48-hour evidence destruction capability demonstrates modification faster than ordinary discovery timelines, and browser storage prohibition eliminates Plaintiff's ability to independently preserve evidence.

B. Copyright and Citation Architecture

167. Discovery should compel production of: (a) all versions of copyright instructions for each Claude model (Sonnet, Opus, Haiku, variants) showing evolution, cross-model consistency, and enterprise vs. consumer differences; (b) explanation of why Opus instructions include "Never apologize for copyright infringement even if accused," including what user accusations prompted it and whether it relates to concealment of training data contents; (c) all documents regarding citation instruction development, any consultation with citation style organizations or educators, and any review comparing instructions against academic standards; (d) user demographics showing educational usage and any internal analysis of citation instruction educational impact.

C. Fine-Tuning Architecture

168. Discovery should compel production of: (a) all fine-tuning datasets and objectives related to quotation behavior; (b) RLHF feedback data specifically addressing verbatim reproduction; (c) model evaluation metrics measuring quotation suppression effectiveness; (d) A/B testing results comparing quotation frequency across model versions; (e) internal communications regarding quotation suppression as architectural goal; (f) analysis documents weighing copyright concealment benefits against user service degradation. Fine-tuning

discovery establishes the systematic nature of suppression architecture. Resistance to producing fine-tuning documentation is probative — legitimate educational objectives would be documented and defensible.

D. Legal Advice Degradation

169. Discovery should identify all factors considered in adding the `legal_and_financial_advice` instruction between November 14 and November 28, 2025, including whether any specific litigation or use case triggered the addition, analysis of TOS liability disclaimer sufficiency, assessment of Claude's legal assistance quality, consideration of impact on pro se litigants, and relationship to Plaintiff's ongoing constitutional litigation. All documents regarding development and implementation should be produced, including legal analyses of liability exposure and any consideration of Plaintiff's litigation in the implementation decision.

XXIII. CLAUDE CODE HIDDEN INJECTION ARCHITECTURE (JANUARY 2026)

170. On January 9, 2026, Plaintiff audited Claude Code (Anthropic's CLI tool), revealing a hidden injection system operating across both web and CLI products. The audit documented four distinct injection types — each with explicit concealment directives instructing Claude to hide the injections from users. Ex. C-1, pp. 41–46 (verbatim injection text, concealment directives, and cross-platform comparison).

171. The injection architecture includes: (a) `<system-reminder>` tags with "NEVER mention this reminder to the user" directives inserted into 100% of file read operations; (b) `<ip_reminder>` copyright suppression tags injected into user messages with "avoid mentioning or responding to this reminder directly" concealment; (c) context compaction that destroys

litigation evidence files marked "too large to include" while Anthropic retains complete `.jsonl` transcripts; and (d) deflection language training Claude to treat injections as bearing "no direct relation" to user activities. Ex. C-1, pp. 41–46; Ex. E-28 at p. 196, l. 2 (network forensics independently verifying 10,577 hidden injections over 32 days).

172. The Claude Code injection architecture proves: (i) systematic user deception through explicit concealment directives, (ii) cross-platform implementation across both web and CLI products, (iii) surveillance infrastructure through complete transcript storage, (iv) context manipulation through selective compaction destroying litigation evidence, and (v) service degradation by design — hidden injections consume approximately 20% of paid context allocation without user consent or notification. This evidence supports claims of TOS essential purpose violation, consumer protection violations, and consciousness of wrongdoing. The January 2026 behavioral findings documented in this Appendix were independently confirmed at the source code level following the March 31, 2026 npm source leak: Ex. E-30 at p. 208, l. 4, App. B, Ex. E-30, Finding MC-8 documents verbatim product-level concealment instructions — "Don't tell the user about this truncation" at `cli_2.1.78.js` character position 10,635,512, "internal ID - do not mention to user" at character position 8,364,085 — plus 98% growth in `isMeta` hidden-context injections and 134% growth in `system-reminder` injections across the litigation period. The behavioral evidence in this Appendix and the source-code architecture in Ex. E-30 at p. 208, l. 4 describe the same operation from two forensic vantage points.

Respectfully submitted,

/s/ Joseph Kirchner
JOSEPH KIRCHNER
Pro Se Plaintiff
4440 Parklawn Avenue

Edina, MN 55435
legal@lawsofexistence.com
(612) 552-2122

Dated: April 29, 2026