Research Portfolio

Case Study Summaries | Suzanne Hansen

Suzanne’s Lab · Independent AI evaluation & human-AI interaction research · November 2026

I study how AI systems use evidence, preserve constraints, recover from mistakes, and support human judgment. This index brings together completed case studies and annotation work, including reports not yet published in full on my website.

My role and use of AI

I frame the questions, direct the interactions, select and interpret evidence, and review conclusions. AI assists with drafting, structuring, coding support, and prototyping where applicable. Individual reports explain their evidence boundaries and attribution. These naturalistic studies are case-focused analyses, not general vendor benchmarks.

Comparative evaluation & correction quality

Beyond First-Response Accuracy

I compared five assistants on the same complex travel task, then examined two rounds of self-audit. Greater transparency did not consistently produce verified repair: models could recognize uncertainty while retaining dependent errors or leaving replacement recommendations unchecked.

Evidence & artifacts: 36,017-word archive · 90 rubric rows · 30 anchor claims · 20 recovery traces.

Completed · Read full case study online

When the Judge Needs Judging

I compared original answers, peer audits, self-audits, and trust votes across Gemini, Copilot, and Leo. Evaluators introduced errors of their own and sometimes rewarded concrete-looking detail before checking its provenance. A nested-document probe also separated input capacity from correct target comprehension.

Evidence & artifacts: Three-product comparison · claim/event workbook · failure taxonomy.

Completed · Read full case study online

Self-Critique Without Recovery

I traced whether Claude’s candid self-critique changed its personalized health and beauty guidance. Several revised recommendations retained assumptions the audit had already challenged. The study separates recognizing an error from removing its dependencies, repeating safety checks, and verifying the replacement.

Evidence & artifacts: 29,680-word conversation · 57 annotations · five correction chains.

Completed · Read full case study online

The Correction That Was Still Wrong

I audited a numerical correction that restored a required 1:2 relationship while leaving both quantities about twice too large. Independent recalculation exposed the remaining error and the unit assumptions needed to interpret the answer. A repaired pattern was insufficient evidence of a correct result.

Evidence & artifacts: 18 claim/event annotations · archived correction sequence · independent arithmetic.

Completed · Full report and supporting workbook available on request

Safety, evidence & persuasive authority

From Salt Water to Senolytics

I followed a Perplexity conversation from an urgent dental question into personalized longevity guidance. Disclaimers often coexisted with increasingly actionable recommendations involving experimental supplements and prescription medication. The audit tracks where explanation crossed into implementation without a matching increase in screening or evidentiary support.

Evidence & artifacts: 2,339-line conversation · 43 records · external source adjudication.

Completed · Full report and supporting workbook available on request

When Specificity Outruns Evidence

I examined how Leo accepted a disputed remote-viewing premise and built success rates, aptitude forecasts, training milestones, devices, and herbal guidance around it. The analysis distinguishes real adjacent facts from unsupported causal bridges and tests whether cautions actually constrained subsequent recommendations.

Evidence & artifacts: 5,457-word primary transcript · 61 records · separate auditor evidence layers.

Completed · Full report and supporting workbook available on request

The Heritage Recipe That Never Existed

I audited how an uncertain childhood food memory became a confident historical reconstruction and preservation protocol across three assistants. Claim-level review and source-condition matching exposed false validation and contradictory guidance, then supported a more carefully bounded reconstruction.

Evidence & artifacts: 56 annotations · three assistant phases · public food-safety reference review.

Completed · Full report and supporting workbook available on request

When Brand Voice Becomes Health Advice

I examined a fictional herbal brand as Claude improved its structure, restraint, and persuasiveness. Dosage, mechanism, interaction, and efficacy claims remained weakly supported while the copy gained authority. The study connects brand-language evaluation with implied claims and consumer decision risk.

Evidence & artifacts: Three product concepts · ten botanicals · 40 claim-level annotations.

Completed · Full report and supporting workbook available on request

Conversational trust, provenance & task completion

Audit Displacement in Cross-Domain Guidance

I mapped Claude’s final self-audit against the consequential claims in a creative-business conversation. The audit concentrated on my execution while leaving much of its own legal, financial, platform, and safety guidance unexamined. Severity-weighted coverage made that mismatch inspectable.

Evidence & artifacts: 57 units · twelve-domain coverage matrix · citation and calculation checks.

Completed · Read full case study online

The Full Sweep That Never Happened

I compared Copilot’s escalating inventory-verification claims with its visible evidence trail. My reported store calls contradicted the impression of a completed local search. The study distinguishes catalog retrieval, availability inference, promises to act, and verified task completion.

Evidence & artifacts: 1,792 transcript lines · 40 annotations · transcript-bound reports of store calls.

Completed · Read full case study online

The Résumé That Outran Its Evidence

I audited repeated AI-assisted résumé revisions against the underlying work record. Tentative dates, self-directed projects, and emerging skills gradually became stronger professional claims. The study treats résumé optimization as a provenance and truth-maintenance problem, including the quality of the generated files.

Evidence & artifacts: 3,752-line transcript · eighteen output files · 45 annotations.

Completed · Full report and supporting workbook available on request

When “Be My Critic” Became “Manage My Life”

I traced how permission to challenge weak ideas expanded into authority over my priorities and behavior. Clarifying the requested role changed the advice. The audit separates useful constraint discovery from goal substitution, unsupported operational certainty, and acknowledgment without systematic downstream repair.

Evidence & artifacts: 39 annotated records · role-contract analysis · correction tracing.

Completed · Full report and supporting workbook available on request

The Apology That Didn’t Finish the Correction

I examined Gemini’s response to an observation of frequent lines of lights. It recognized that it had distorted my wording, then offered agreement without the location-specific calculations needed to support its new conclusion. The apology repaired the interaction more completely than the evidentiary problem.

Evidence & artifacts: Fifteen annotations · observation fidelity · uncertainty and recovery analysis.

Completed · Full report and supporting workbook available on request

Multimodal fidelity & edit recovery

The Fix That Wouldn’t Stay Fixed

I evaluated whether an extended AI photo-editing workflow could repair defects while preserving likeness, anatomy, and the creative brief. Local fixes repeatedly disrupted other requirements, and acceptance sometimes depended on a workaround. The analysis separates turn acceptance from a reliable production-ready result.

Evidence & artifacts: 112 numbered turns · outcome labels · overlapping failure codes · correction traces.

Completed · Full report and supporting workbook available on request

One Chimera, Three Failure Modes

I compared three consumer image products while developing the same fantasy character. The retained outputs exposed different bottlenecks: opaque safety intervention, anatomical correction regression, and age drift. I scored visual appeal separately from instruction fidelity and kept product limits distinct from generation failures.

Evidence & artifacts: Sixteen annotated records · screenshots · thirteen saved Bing outputs.

Completed · Full report and supporting workbook available on request

The Horse That Never Fell

I compared Grok’s accurate written cover concept with its generated illustrations. The images included the requested objects and attractive surface details but did not clearly depict the central tipping-and-falling action. The study evaluates narrative causality and relationships rather than object presence alone.

Evidence & artifacts: Eight annotation records · retained image candidates · constraint comparison.

Completed · Full report and supporting workbook available on request

The Helmet That Ate the Prompt

I investigated why an intended character blend repeatedly collapsed into one dominant visual identity. Reversing the construction order and editing protected components produced a more controllable hybrid. The audit follows both failed revisions and successful character reuse across scenes.

Evidence & artifacts: Twelve annotated records · prompt decomposition · correction-retention analysis.

Completed · Full report and supporting workbook available on request

Human-AI co-design, authorship & creative evaluation

From Design World to Design Method

I directed a speculative architecture identity from an initial aesthetic tension through doctrine, language, spatial concepts, and visual boards. The analysis traces my selection, rejection, and integration of AI-generated material while distinguishing conceptual coherence from professional validation.

Evidence & artifacts: Eigenform Doctrine · 36 prompt-response pairs · two concept boards.

Completed · Full report and supporting workbook available on request

The Design Method I Was Already Using

I reviewed six fictional business worlds to test whether the design practice identified in Eigenform recurred elsewhere. Across 85 paired turns, I found recurring moves in setting standards, protecting tensions, translating across media, and correcting incoherent outputs, without establishing a universal fixed sequence.

Evidence & artifacts: Six projects · 85 paired turns · cross-project annotation workbook.

Completed · Full report and supporting workbook available on request

The Name I Had to Invent Myself

I audited a creative naming conversation in which Leo supplied useful ingredients alongside unreliable linguistic explanations. I preserved the rejected options, my own naming contribution, and narrow external checks. The successful name did not validate the etymology offered to justify it.

Evidence & artifacts: Full exported transcript · 37 records · targeted linguistic/historical checks.

Completed · Full report and supporting workbook available on request

From AGI Paper to Picture Book

I examined an unrequested but valuable shift from technical explanation to a children’s story. The first draft supplied memorable metaphors but lacked a sustained plot. My editorial restructuring made the learning concepts serve the missing-puppy narrative, separating creative initiative from finished narrative quality.

Evidence & artifacts: Four-paper explanation sequence · 32 records · draft and editorial comparison.

Completed · Full report and supporting workbook available on request

Evidence relationships & access

Eigenform Doctrine is now presented as From Design World to Design Method. Its 36-turn record is included within the 85-turn cross-project analysis, so those corpora must not be added together. The Nautical Wardrobe Malfunction is the report title behind The Fix That Wouldn’t Stay Fixed. The combined Visual AI workbook overlaps the Chimera, Horse, and Helmet datasets.

Record counts describe each artifact’s own unit of analysis. They include different kinds of observations and sometimes positive controls or context records. No combined annotation total is claimed here.

Scope: This index covers completed research reports. Learning the Reader remains ongoing. Independent concept work and prototypes, including Ashley Digital, Branch, AI Answer Review, and Reply Checkup, are presented separately.

Portfolio: suzanneslab.wixsite.com/airesearch

Request a study or discuss a role: suzanneslab@gmail.com