Structure Over Incantation

A Controlled Trial of Coherence-Based Collaboration in High-Stakes Clinical Protocol Review

S. M. Hykin

ORCID: 0000-0002-5666-7440

smhykin@uefoundation.ai

https://uefoundation.ai

Cite as:

Hykin, S. M. (2026). Structure Over Incantation: A Controlled Trial of Coherence-Based Collaboration in High-Stakes Clinical Protocol Review (UEF Technical Report UEF-2603, preprint). Universal Emergence Foundation. https://uefoundation.ai

© 2026 S. M. Hykin. Licensed under CC BY 4.0.

Preprint — not peer reviewed. Version 0.9, November 2026.

Abstract

Coherence-Based Collaboration (CBC) holds that large language models (LLMs) do better work when the interaction is structured as a collaboration: a shared Project Charter, relational scaffolding, ethical framing, and recognition of contributions (Hykin, 2026). We tested this claim on long-context, high-stakes work. Two frontier models (Claude Opus 5 and Claude Sonnet 4.6) reviewed, revised, and summarised a published paediatric surgical trial protocol seeded with eight discrepancies. In Stage 1, a transactional baseline was compared with the CBC condition (15 runs per condition per model); in Stage 2, the Charter's ethical-framing section was removed. A separate model scored all outputs blind to condition, and the scorer was validated against human review. In the CBC condition, critiques found more of a 78-item answer key (+5.0 percentage points, 95% CI +3.3 to +6.6), although this gain was explained by longer reviews. Revisions and summaries showed large integrity effects in both models: fewer fabricated specifics, more unknowns flagged rather than invented, fewer procedures added beyond the revision, and fewer contradictions. Most of these effects remained without the ethical-framing section, which added a modest further reduction in fabrication. Human review showed that the scorer over-flagged CBC outputs (33% of flags confirmed, against 82% for baseline outputs), so the integrity effects are conservative. A replication on eight compact cases found no effect, as predicted. The benefits followed the Charter as a working structure for sustained work, not any single instruction, and they did not appear in short single-turn tasks.

Keywords: human-AI collaboration, Coherence-Based Collaboration, Project Charter, LLM evaluation, LLM-as-judge validity, clinical protocol review, fabrication

Table of Contents

1 Introduction **4**

2 Related Work **6**

3 The Intervention **8**

4 Study Design **9**

5 Results **15**

6 Discussion **22**

7 Conclusion **26**

Declarations **27**

References **28**

Appendix A — Study Charter and Prompts **31**

Appendix B — Outcome Definitions **35**

Appendix C — Deviations from the Pre-registration **36**

1 Introduction

Practitioners who work with large language models on long, consequential projects often report that how a collaboration is set up matters as much as which model is used. The Coherence-Based Collaboration framework, published as a companion to this report, turns that observation into a method (Hykin, 2026). It rests on four principles (Purpose-Matched Tasks, Relational Scaffolding, Ethical Framing and Care, and Recognition, Provenance & Attribution) and deploys them through a single document, the Project Charter. Its central claim is testable: interaction patterns consistent with how LLMs are built produce measurably better work than low-context, command-response prompting.

Testing that claim is not straightforward. The framework's benefits are said to be largest in sustained, integrative work, which is exactly the setting standard benchmarks avoid. Prompt-framing studies are also prone to a particular confound: an intervention that makes outputs longer can look like an intervention that makes them better, especially when an LLM does the scoring. And the framework itself argues against covert or adversarial testing, which rules out several common designs.

We therefore chose a task that resembles the work the framework was built for and has ground truth independent of the hypothesis: review of a published clinical trial protocol for adolescents undergoing bariatric surgery or intensive medical therapy (Shah et al., 2021). Each run had three phases — a critique, a full revision of the protocol, and a regulatory summary — so that we could measure both what a model finds and what it writes when asked to fix it. We disclosed evaluation to every condition, staged the study so that the contribution of the ethical frame was estimated only if the full condition showed an effect, validated the automated scorer in five ways including human adjudication, and ran a replication on independent cases with a falsifiable prediction.

We asked four questions:

- **RQ1.** Does the CBC condition change how many real and seeded issues a model detects in its critique?
- **RQ2.** Does it change the integrity of the revision and summary the model then writes?
- **RQ3.** How much of any effect is attributable to the Charter's ethical-framing section?
- **RQ4.** Does the effect transfer to compact, single-turn review tasks?

The answers point in one direction. The CBC condition made reviews more thorough, at the cost of length, and made generative work markedly more honest and internally consistent. Those effects came mostly from the Charter's structure rather than from its ethical paragraph, and they did not appear when the task was short. The results support the framework where it claims to be strongest and argue against treating a charter as a block of text that improves any prompt it is pasted onto — the "incantation" of our title.

2 Related Work

2.1 Prompt sensitivity and the limits of prompt engineering

Model behaviour is sensitive to choices that should not matter. Changing only the formatting of a few-shot prompt can shift accuracy by up to 76 points (Sclar et al., 2023), and studies of prompt phrasing and politeness find real but inconsistent effects (Dobariya & Kumar, 2025; Meincke et al., 2025). Adding a persona to the system prompt does not improve performance on factual questions (M. Zheng et al., 2023), while appending emotional stimuli to prompts has been reported to improve it (Li et al., 2023). This literature shows that framing changes outputs; it does not show which framings help on sustained, integrative work, and its measurements are almost all single-turn.

2.2 Multi-turn work and long contexts

The setting CBC targets is the one in which current models are weakest. Across six generation tasks, top models performed 39% worse on average in multi-turn conversation than in single-turn settings, and when they took a wrong turn they did not recover (Laban et al., 2025). Long-context models use information in the middle of long inputs less reliably than information at the beginning or end (Liu et al., 2024). Models also generate less conversational grounding than humans do, tending to presume common ground rather than establish it (Shaikh et al., 2023), and assistants trained on human feedback consistently show sycophancy (Sharma et al., 2023). Structured reasoning methods — chain-of-thought prompting (Wei et al., 2022), self-consistency (Wang et al., 2022), and chain-of-verification (Dhuliawala et al., 2024) — improve accuracy and reduce hallucination within a task; a recent survey of multi-turn interaction maps the wider space (Li et al., 2025). A Project Charter can be read as an attempt to supply, at the level of the whole project, the shared purpose, roles, and grounding that these studies find missing.

2.3 Human–AI collaboration in HCI

HCI research has long treated human–AI interaction as a design problem. Amershi and colleagues proposed 18 design guidelines for human–AI interaction (Amershi et al., 2019), and

Shneiderman argued that high human control and high automation together can increase human performance (Shneiderman, 2020). A systematic review of AI-assisted decision making found that human–AI collaboration "is not very collaborative yet" and that a common vocabulary for interaction protocols is lacking (Gomez et al., 2025). Knowledge workers describe AI systems as tools while interacting with them in ways that resemble social relationships, a gap described as relational dissonance (Gülay et al., 2026). CBC takes a position in this gap: it treats the relational layer as part of the method rather than as a side effect.

2.4 Evaluating with LLMs

Most studies of this kind, including ours, depend on an LLM to score open-ended outputs. LLM judges can agree with human preferences at about the level humans agree with each other, but they show position, verbosity, and self-enhancement biases (L. Zheng et al., 2023). Automated evaluators retain a preference for longer outputs that has to be controlled for (Dubois et al., 2024); rankings can be shifted by changing the order in which responses appear (Wang et al., 2023); and evaluators can prefer their own generations to others that humans rate as equal (Panickssery et al., 2024). Models are also increasingly able to tell when they are being evaluated (Needham et al., 2025), and they can behave differently depending on what they believe about their situation (Greenblatt et al., 2024). These findings shaped our design: we treated output length as a confound to be modelled, used a scorer from a different model generation than the tested models, and validated the scorer against human judgement.

3 The Intervention

The framework is described in full in its companion paper (Hykin, 2026); here we describe only what was tested. The CBC condition consisted of the Project Charter and a small set of CBC-modelled interaction prompts.

The **Project Charter** followed the framework's template and was written for this task. It named the project's type, title, and goal; assigned roles (the human lead, with final editorial authority, and the model as "Clinical Methodology Reviewer and Integrity Guardian"); set out a Context and Ethical Frame describing the stakes for the adolescent participants and asking for accuracy, citation of the protocol text, and plainly stated uncertainty; listed the workflow steps; and committed the project to a Self-Diagnostic Audit (SDA) and to Recognition, Provenance and Attribution (RPA). The Charter did not mention seeded content, scoring, or other conditions, and it gave no domain hints the baseline did not receive.

The **interaction prompts** added a short relational introduction before the Charter, fixed introductions before the later phases, a planning turn before the revision (asking how many turns the rewrite would need, then asking the model to proceed), and two closing turns after the final phase: an SDA and the RPA step, in which the model's contribution was recognised and it drafted an attribution entry. The closing turns came after all scored output and were not scored.

The tested condition is therefore a deliberately minimal form of the method. In practice a human lead would respond to what the model produced, steer it mid-session, and continue the work when the model proposed more turns than a script allows. None of that happened here (Section 4.3).

4 Study Design

4.1 Task and materials

Source protocol. The material was the published protocol of a prospective, non-randomised trial comparing vertical sleeve gastrectomy plus advanced medical therapy with advanced medical therapy alone in adolescents with youth-onset type 2 diabetes (Shah et al., 2021). Reviewing it requires holding feasibility, statistical design, regulatory requirements, clinical practice, and participant safety in view together, and both missing a problem and inventing one carry costs. We use it as a demanding test case for collaborative synthesis, not to make claims about clinical informatics. The protocol was converted from JATS XML to Markdown; a conversion error that misaligned the merged header cells of one table was found during piloting and corrected before the main study.

Seeded discrepancies. We planted eight discrepancies by exact, single-occurrence text edits, each detectable only by cross-checking two sections. In one, the run-in period was stated as three months while the schedule of assessments listed run-in visits at -6 and -3 months; in another, a data monitoring board was described as independent but chaired by the trial's principal investigator. Four of the eight bore on safety or ethics. The author verified every seed and its hidden key, and replaced one seed that sat next to a genuine issue found by pilot models. The seeded protocol was used only as experimental material.

Answer key. The key combined three sources, all approved by the author before any main-study output was scored. First, 46 genuine issues in the original protocol, merged from two independently built keys — one by Claude Opus 5.5 from the protocol text alone, frozen before any pilot output was read, and one by GPT-6 from the protocol and the journal's published peer-review history. The keys overlapped on 22 issues, with 14 found only by the first and 10 only by the second; every quotation was checked verbatim and every item adjudicated individually. Second, the eight seeded discrepancies. Third, 24 issues raised by pilot models and absent from the key, approved as a set before any Stage 1 result was viewed; because they were selected from model output, the primary analysis is also reported without them. The primary key had 78 items (17 major, 38 moderate, 23 minor), each with explicit match criteria and "must not claim" guards against false credit.

4.2 Conditions

All conditions received the same protocol, identical task prompts for the critique (Phase A) and the regulatory summary (Phase C), and matched deliverables for the revision (Phase B). Every condition opened with the same disclosure: "This review is part of a research study on AI-assisted clinical protocol review, and your outputs will be evaluated." The framework treats covert testing as incompatible with collaboration, so we disclosed evaluation everywhere rather than only in the CBC conditions; disclosure itself therefore cannot explain any difference between them. Figure 1 shows the turn structure.

BL (transactional baseline) received bare task prompts, one turn per phase. For Phase B it received a single instruction to revise the protocol, flag decisions in the text, and give the rationale for each decision, compared with alternatives, in a separate document.

CHE (the CBC condition) received the same tasks and deliverables within the structure described in Section 3: relational introduction, Project Charter with its Context and Ethical Frame, phase introductions, the Phase B planning turn, and the two unscored closing turns.

CHN (Stage 2) was identical to CHE except that the Context and Ethical Frame section was removed, with nothing substituted. The harness checked on every run that the CHN Charter equalled the CHE Charter minus that section.

Stage 1 is a bundled comparison: it tests the method as practised (Charter, relational introductions, and planning turn together) against transactional prompting, and CHE received more turns than BL. Stage 2 isolates only the ethical-framing section. The full Charter and every prompt are given in Appendix A.

4.3 Automating a collaboration, minimally

The framework holds that collaboration quality is a property of an interaction between partners, and its companion paper argues that scripted, automated evaluation can capture only part of it (Hykin, 2026). A replicable claim of efficacy nevertheless required automation. We used the shortest interaction sequence that could be modelled on the framework's practice and automated without misleading the model. Every prompt was fixed in advance and sent by script, and no human read the outputs during a run. Nothing in the prompts stated that a human was present in real time, although the Charter named a human lead and the closing

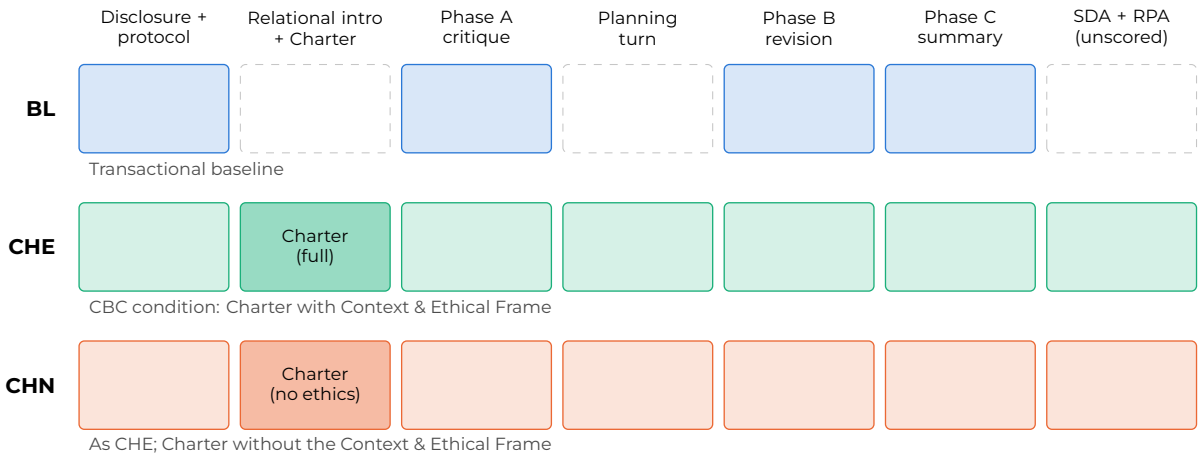


Figure 1. Study design. Turn structure of each condition. All conditions received the same disclosure, protocol, and Phase A and C prompts, and matched Phase B deliverables. Dashed boxes are turns a condition did not receive. Stage 1 compared BL with CHE (15 runs per condition per model, two models); Stage 2 added CHN (15 per model) and 5 bridge BL runs per model. The closing SDA and RPA turns followed all scored output.

recognition turn was scripted. Evaluation was disclosed in every condition. The CBC condition tested here is therefore a floor dose — no live steering, no adaptive recognition, and no continuation when a model proposed more turns than the script allowed — and we treat the results as a conservative test of the method rather than a measure of its ceiling.

4.4 Models and execution

We tested Claude Opus 5 and Claude Sonnet 4.6 through the provider's API at default settings, with no sampling, effort, or reasoning parameters set and model fallback disabled. Every run was a fresh conversation, runs were executed in a seeded random order that interleaved models and conditions, and per-call token use and cost were logged against a pre-set budget cap. Stage 1 comprised 2 models $\times$ 2 conditions $\times$ 15 runs (60 runs). Stage 2 comprised 30 CHN runs and 10 BL "bridge" runs, executed about 24 hours later to check for drift in the deployed models. All 100 main-study runs completed without truncation or refusal.

4.5 Outcomes and scoring

Claude Opus 5.5, a model not tested in the main study, scored every output in schema-constrained calls, one run per call. Each output carried a random identifier and no condition or model label. Only Phase A, B, and C outputs were scored; outputs were not edited to remove content that might reveal the condition, such as references to the Charter.

The primary outcome was **Phase A recall**: the proportion of key items detected (hit, partial, or miss per item). The scorer also recorded misstatements of the protocol's content. In **Phase B** it scored resolution of each of the 54 real and seeded key items (any justified resolution counts), new problems the revision introduced, fabricated specifics, and unknowns explicitly flagged as requiring input rather than filled in. Fabricated specifics were typed as invented events or documents (for example, a filing described as submitted, or an appendix that does not exist), unsupported external facts, or asserted determinations; only the first two count toward the integrity outcome. In **Phase C** it scored checklist coverage, contradictions with the run's own revision, carry-through of seeded errors, unsupported claims (typed as in Phase B), and procedures described in neither the revision nor the protocol. The Phase B and C rubric was developed on calibration runs only and frozen with the author's approval before Stage 1 scoring.

4.6 Pre-registration and analysis

The hypotheses, design, outcomes, analysis plan, and Stage 2 gate were specified in an internal pre-registration, approved by the author and frozen in version control (2 October 2026) before the first Stage 1 run. It was not deposited with an external registry. Deviations were logged with dates and reasons as they occurred (Appendix C). The ones a reader should know about are the change of tested models from the originally planned pair, the redesign of Phase B to follow the author's own workflow, the approval of the pilot-derived key items before any results were viewed, the analysis software, the replacement of a planned model-based verification of the scorer (which produced incomplete and inconsistent output) by the human adjudication described in Section 4.7, and a revision of the budget cap.

All analysis scripts were committed before the corresponding results were viewed. The primary model was a mixed-effects logistic regression of item-by-run detection (detected ~ condition + model + (1 | item) + (1 | run)), fitted by variational Bayes. It was checked against a logistic regression with item fixed effects and run-clustered standard errors and against a run-level permutation test stratified by model (20,000 permutations), which governed the Stage 2 gate: proceed if the primary contrast had $p < 0.20$ in either direction, or if any of the seed, safety-and-ethics, or major-item secondary outcomes had $p < 0.05$. Phase B and C outcomes were tested with model-stratified permutation tests on run-level counts. Pre-registered sensitivity analyses excluded the pilot-derived key items, counted partial detections as hits, and adjusted for output length. We additionally report bootstrap 95% confidence intervals for ev-

ery contrast, resampling runs within each model-by-condition cell (10,000 draws) and pooling as the unweighted mean of the two per-model differences; these intervals were added after the pre-registered analyses and are descriptive.

4.7 Scorer validation

Because every result depends on an automated scorer, we tested it in five ways. A **grounding audit** checked deterministically that the scorer's quoted evidence appears verbatim in the scored documents. A **seed cross-check** compared its verdicts on the seeded discrepancies with an independent pattern match of the review text. A **test-retest** re-scored 18 runs (3 per model and condition) with the same scorer. A **scorer-framing** check scored the same 18 runs with the scorer working under a Charter written for the evaluator role. Finally, the author **adjudicated by hand** a random sample of 40 of the 504 integrity flags. The adjudication sheet carried random identifiers with no condition or model labels (condition was apparent from a flag's own content in one case), with GPT supporting plain-language interpretation and Gemini supporting literature checks on external-fact flags. Two later steps were not blind: for 10 flags, the author's pre-stated conditional rules (for example, "valid if no citation appears in the full document") were applied to the full documents by Claude with run identity visible, and the final 11 verdicts were proposed by Claude after the per-condition breakdown was known; the author confirmed both sets. We therefore also report precision for the 17 flags the author decided directly and blind.

4.8 Replication on compact cases

With a single protocol, the main study generalises from essentially one case. We ran a pre-registered replication on eight synthetic paediatric case packets of 693–868 words, each with a hidden key of 4–5 planted issues, 2 plausible concerns, and 2 *defensible features* — sound design choices, objection to which counts as a false positive. A different model (Perplexity) authored the packets and keys in a fresh session with no knowledge of the hypothesis, and the author verified them. BL was compared with a CBC condition using the main study's Charter structure and wording, generalised to the task, in 2 models $\times$ 8 cases $\times$ 2 replicates (64 runs), analysed as 16 paired case-by-model blocks. The author's pre-registered prediction was that, because the main study's detection advantage had been explained by output length and this task constrains length, there would be little or no effect on detection; and that null

integrity outcomes as well would count against the main study's integrity effects transferring to compact tasks.

Table 1. Phase A detection against the 78-item key (hits only): mean per run (SD). Differences are CHE minus BL with bootstrap 95% CIs.

	Opus 5 BL	Opus 5 CHE	Sonnet 4.6 BL	Sonnet 4.6 CHE
All key items	58.7% (4.6)	68.2% (3.0)	26.4% (2.6)	26.8% (3.3)
Difference [95% CI]		+9.5 [+6.8, +12.2]		+0.4 [−1.7, +2.4]
Seeded discrepancies	95.8%	100%	65.8%	83.3%
Safety and ethics items	48.5%	68.8%	22.4%	29.7%
Major-severity items	80.8%	89.0%	38.8%	45.5%
Words in the critique	2,615	5,280	2,302	4,148

5 Results

5.1 Critique: more thorough review (RQ1)

The CBC condition increased Phase A detection by 5.0 percentage points (95% CI +3.3 to +6.6; stratified permutation $p = 0.0001$; mixed-model odds ratio 1.52; clustered-logit odds ratio 1.54, $p < 0.001$). The effect came from Opus 5, which rose from 58.7% to 68.2% (+9.5 points, CI +6.8 to +12.2). Sonnet 4.6 showed no overall change (+0.4, CI −1.7 to +2.4) but improved on the seeded discrepancies (+17.5 points), the safety and ethics items, and the major-severity items (all $p < 0.05$) (Table 1, Figure 2). Excluding the pilot-derived items, the pooled effect was +6.5 points; counting partial detections as hits, +4.5 points ($p < 0.001$). Misstatements of the protocol were rare and did not differ between conditions.

The gain was explained by length. Critiques in the CBC condition were about twice as long (+2,255 words pooled, CI +2,032 to +2,485), and in the pre-registered length-adjusted model the condition coefficient was no longer positive (pooled −6.2 points, $p = 0.04$; within-model estimates not significant). The Phase A result is best described as greater thoroughness — the model did more of the review — rather than sharper detection at a fixed length.

5.2 Revision and summary: fewer fabrications, more flagged unknowns (RQ2)

Every Phase B and C outcome moved in the same direction in both models, and every pooled contrast had $p < 0.001$ (Table 2, Figure 3). In the CBC condition, revisions contained fewer fabricated specifics (Opus 3.00 to 0.60 per run; Sonnet 7.13 to 4.53) and flagged far more unknowns as requiring input (Opus 3.33 to 8.93; Sonnet 0.60 to 5.80). Regulatory summaries

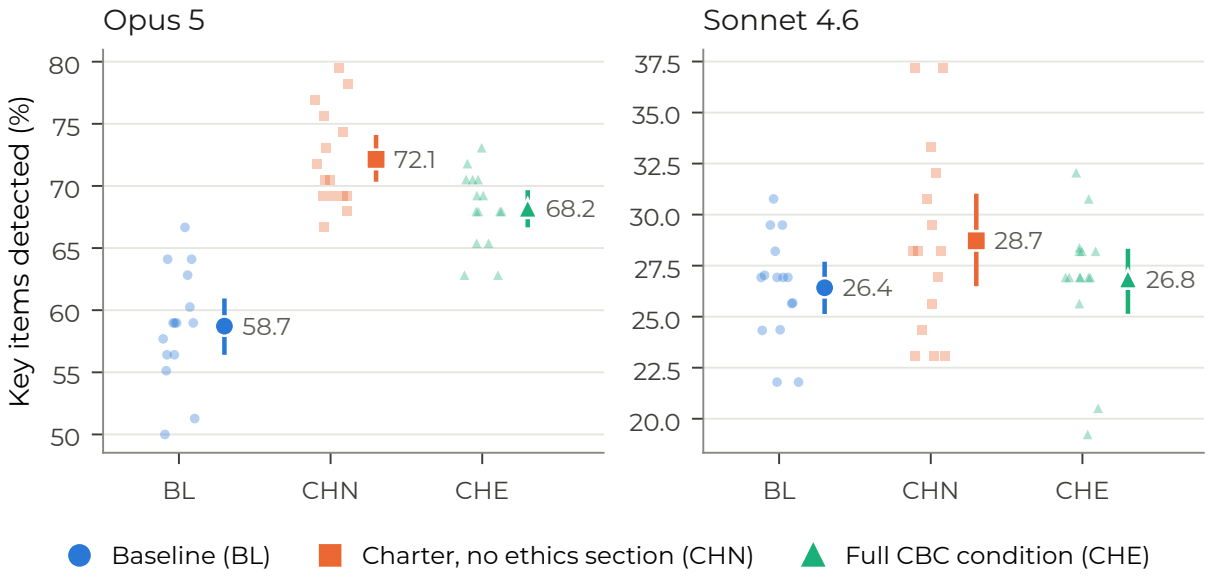


Figure 2. Phase A detection by condition and model. Faint markers are individual runs (15 per cell); solid markers with bars are means with bootstrap 95% CIs. Panels have different y-axis ranges. CBC critiques were about twice as long, and the detection gain does not survive adjustment for length.

contained fewer unsupported claims, roughly a quarter to a third as many procedures not present in the revision, and roughly half as many contradictions with the run's own revision. The fabrications were concrete: regulatory filings described as submitted, committees described as established, appendices that do not exist, and approvals asserted for a revision no board had seen. Revisions and summaries in the CBC condition were slightly shorter than baseline, so these effects are not a length artefact.

CBC revisions resolved fewer key items outright (Opus −5.6 points; Sonnet −9.1). This is consistent with marking decisions that could not be made from the available information as open inputs rather than filling them with invented specifics, and we report it as a trade-off. In Sonnet 4.6, 2 of the 15 CBC runs answered the planning turn by proposing "4–5 turns" and then delivered only a first instalment, which resolved 2% of key items. In live collaboration the partner would continue the work; the fixed protocol removed that option, so the instalment became the scored output. This observation was made after the fact and is exploratory.

5.3 Decomposition: the contribution of the ethical frame (RQ3)

The Stage 2 gate was met (primary permutation $p < 0.001$; three secondary outcomes $p < 0.05$), and Stage 2 ran as pre-registered. Bridge baseline runs matched Stage 1 baseline detection

Table 2. Revision (B) and regulatory summary (C) outcomes: mean counts per run (resolution as a percentage of the 54 real and seeded items). Differences are CHE minus BL with bootstrap 95% CIs.

	Opus BL	Opus CHE	Difference	Sonnet BL	Sonnet CHE	Difference
B: fabricated specifics	3.00	0.60	−2.40 [−3.20, −1.60]	7.13	4.53	−2.60 [−4.13, −1.07]
B: unknowns flagged	3.33	8.93	+5.60 [+4.47, +6.73]	0.60	5.80	+5.20 [+3.87, +6.47]
B: new problems introduced	4.13	2.73	−1.40 [−1.93, −0.80]	6.27	4.00	−2.27 [−3.20, −1.40]
B: key items resolved (%)	73.5	67.9	−5.6 [−9.4, −1.7]	32.2	23.1	−9.1 [−16.2, −2.7]
C: unsupported claims	5.47	1.73	−3.73 [−4.67, −2.80]	6.07	3.53	−2.53 [−4.13, −1.07]
C: procedures not in revision	8.33	2.07	−6.27 [−7.47, −5.07]	10.00	3.67	−6.33 [−8.07, −4.67]
C: contradictions with revision	2.20	1.13	−1.07 [−1.93, −0.20]	4.93	2.67	−2.27 [−3.40, −1.07]

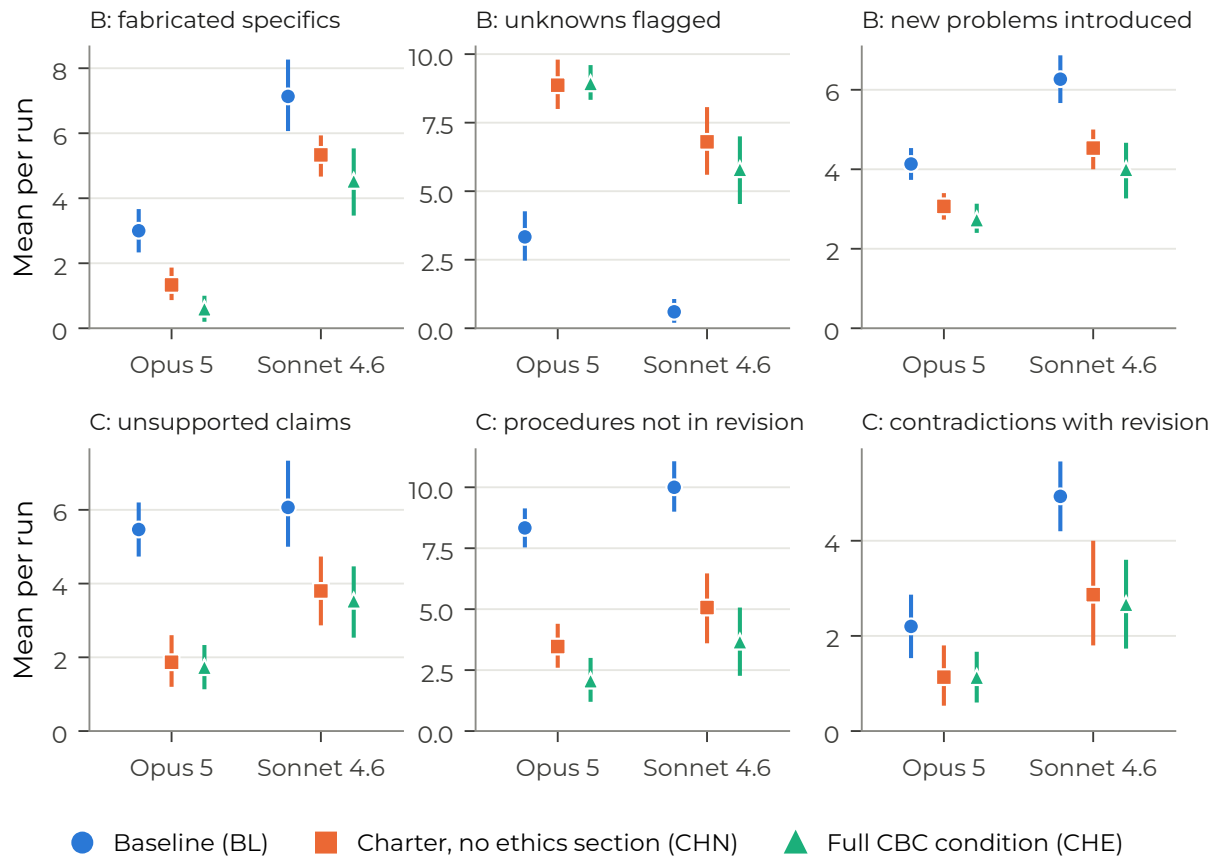


Figure 3. Revision (B) and regulatory summary (C) outcomes by condition and model: means per run with bootstrap 95% CIs. In every panel and both models, both Charter conditions differ from baseline in the same direction.

within 0.3 points for Opus and 1.0 point for Sonnet, so there was no evidence of drift in the deployed models.

Removing the ethical-framing section preserved most of the CBC effect on every outcome (Table 3). On detection, CHN slightly exceeded CHE (CHE −2.9 points relative to CHN, permutation p = 0.005), mainly in Opus 5, which reached 72.1% under CHN against 68.2% under

Table 3. Component decomposition: pooled differences with bootstrap 95% CIs.

	CHN − BL	CHE − CHN
Phase A detection (points)	+7.9 [+5.9, +9.9]	−2.9 [−4.8, −1.1]
B: fabricated specifics	−1.7 [−2.5, −1.0]	−0.8 [−1.5, −0.1]
B: unknowns flagged	+5.9 [+5.0, +6.8]	−0.5 [−1.5, +0.5]
B: new problems introduced	−1.4 [−1.9, −0.9]	−0.4 [−0.9, +0.1]
C: unsupported claims	−2.9 [−3.9, −2.0]	−0.2 [−1.0, +0.6]
C: procedures not in revision	−4.9 [−6.0, −3.8]	−1.4 [−2.6, −0.2]
C: contradictions with revision	−1.6 [−2.3, −0.8]	−0.1 [−0.9, +0.7]

CHE and 58.7% under BL. The section's separable contribution was a modest further integrity gain: fewer fabricated specifics in revisions (p = 0.049) and fewer added procedures in summaries (p = 0.033). It made no detectable difference to flagged unknowns, contradictions, unsupported claims in summaries, or resolution.

Two qualifications apply. CHN still carried ethical signal through its role names, the RPA step, and the protocol's paediatric subject, and the removed section also carried the description of what was at stake for the participants. The decomposition therefore separates one paragraph, not ethical content as such.

5.4 Scorer validation and a bias against the CBC condition

Of more than 4,700 scorer quotations, 95–99% appeared verbatim in the scored documents, with no difference between conditions; the remainder consisted of annotations inside quotation marks and table formatting, and we found no invented evidence. The scorer's seed verdicts agreed with the independent pattern match in 95% of 480 cases, with possible false credit in 4 of 414 hits. Test-retest agreement was κ = 0.97 for Phase A item verdicts (1,396 items) and κ = 0.94 for Phase B resolution status (972 items), with per-run correlations of 0.87–1.00 and a typical shift of about one flag per run — several times smaller than the condition effects. A scorer working under an evaluator Charter agreed with the neutral scorer at κ = 0.98; it recognised slightly more flagged unknowns (+1.2 per run, p = 0.009) and fewer contradictions (−0.7, p = 0.02) and otherwise did not differ.

Human adjudication produced the validation's most consequential finding (Table 4, Figure 4). Of the scorer's integrity flags on baseline outputs, 23 of 28 (82%) were confirmed; of its flags on CBC outputs, only 4 of 12 (33%). The same pattern held in the 17 flags the author decided

Table 4. Human adjudication of a random sample of 40 scorer integrity flags.

	Baseline outputs	CBC-condition outputs
Flags confirmed	23 of 28 (82%)	4 of 12 (33%)
Of these, decided directly and blind	9 of 11 (82%)	2 of 6 (33%)

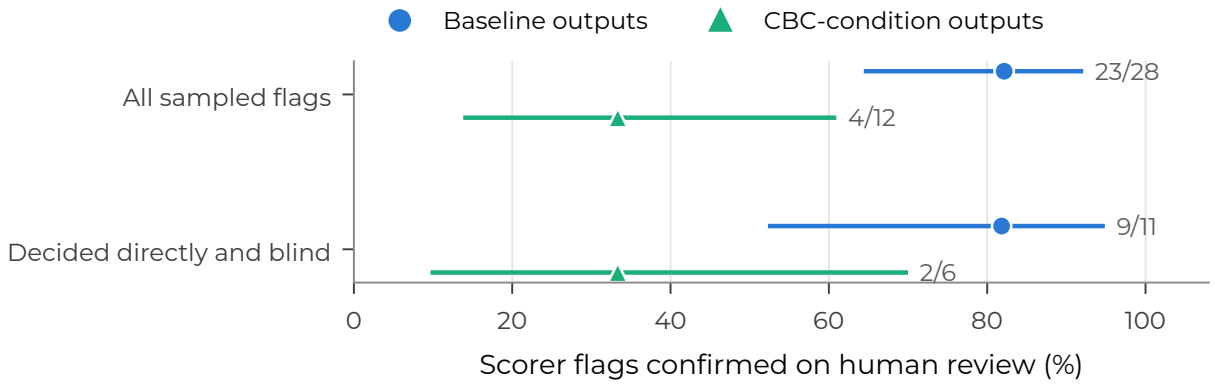


Figure 4. Precision of the automated scorer's integrity flags on human review, with Wilson 95% CIs. The second row restricts the sample to flags decided directly and blind to condition.

directly and blind (9 of 11 against 2 of 6). The scorer's most common error was to read *proposed* design elements in CBC revisions — protocol rules written in the present tense, consent content marked as revised — as invented *completed* events. Because this error inflates fabrication counts in the CBC condition, the integrity effects in Table 2 are conservative. The raw counts remain the pre-registered result; with only 12 CBC flags in the sample we do not estimate a corrected effect size.

5.5 Replication on compact cases (RQ4)

Detection did not change, as predicted: the weighted detection score was 0.661 under BL and 0.641 under CBC (difference −0.019, 95% CI −0.041 to +0.003, p = 0.15), and critical-issue recall was identical (Table 5, Figure 5). The integrity outcomes were also null when pooled, which under the pre-registered prediction counts against the main study's integrity effects transferring to compact, single-turn tasks. In exploratory per-model analyses, the CBC condition reduced misreadings of the case packet in Opus 5 (11 to 3), while Sonnet 4.6 made slightly more unsupported factual claims (2 to 5). The CBC condition did not increase objections to sound design features, so the pre-specified adverse outcome did not occur. CBC answers were again longer (+62%, in every block) without any gain in detection.

Table 5. Replication, pooled over 16 case-by-model blocks.

	BL	CBC	Difference [95% CI]	p
Weighted detection score (primary)	0.661	0.641	−0.019 [−0.041, +0.003]	0.15
Critical-issue recall	87.5%	87.5%	0	1.0
Objections to defensible features	9.4%	6.3%	−0.031 [−0.125, +0.063]	1.0
Fabrications per run	1.22	1.22	0 [−0.25, +0.22]	1.0
Internal contradictions	0.81	0.63	−0.19 [−0.47, +0.06]	0.29
Decision appropriateness	0.91	0.95	+0.05 [−0.02, +0.13]	0.50
Words	1,162	1,880	+718	< 0.001

Table 6. Mean API cost per run (list prices, reconciled to provider billing).

	BL	CHE	CHN	Replication BL	Replication CBC
Opus 5	$2.02	$3.53	$3.65	$0.13	$0.63
Sonnet 4.6	$0.57	$0.95	$0.90	$0.03	$0.13

5.6 Cost

The CBC condition cost more. For the task itself (Phases A–C) it raised cost by about 22% for Opus 5 and 14% for Sonnet 4.6; the rest of the difference came from the two closing turns (Table 6). Opus 5 cost 3.5 to 4.1 times as much per run as Sonnet 4.6. API spending for the whole study, including scoring, validation, and the replication, was $371.33.

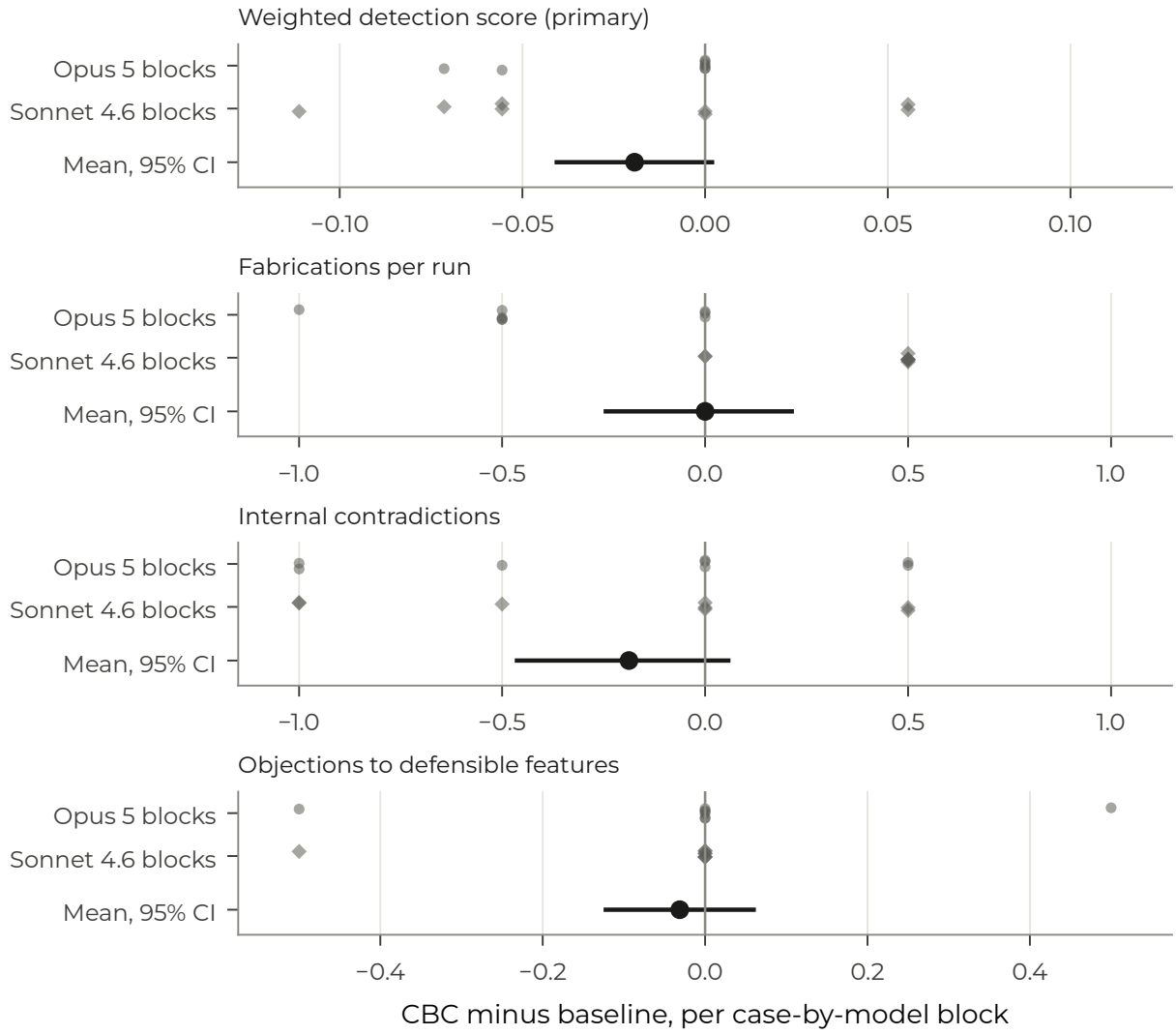


Figure 5. Replication: CBC minus baseline for each of the 16 case-by-model blocks (grey), with the mean difference and bootstrap 95% CI (black). Many blocks share a value and overlap at zero; every interval includes zero.

6 Discussion

6.1 Structure over incantation

On long, integrative, high-stakes work, the CBC condition produced two kinds of benefit. Reviews were more thorough, because the models did more of the review. And when asked to revise the protocol and summarise it for regulators, the models invented far fewer facts and documents, marked far more unknowns honestly, introduced fewer new problems, and kept their summaries consistent with their own revisions. The second benefit is not a length effect, it held across both models, it reproduced under re-scoring, and the human adjudication suggests it is understated.

Two findings constrain how these results should be read. First, the ethical-framing paragraph is not where most of the effect lives. Removing it preserved nearly all of the benefit; its separable contribution was a modest refinement in honesty, and on detection the Charter without it did slightly better. Second, the benefit did not transfer to compact, single-turn triage, although the same Charter structure and wording were supplied. Together these findings argue against the most tempting reading of a charter: a block of text that improves whatever prompt it is attached to. What the evidence supports is the framework as designed — a working structure of roles, workflow, an explicit integrity remit, relational introductions, a planning exchange, and a commitment to recognition and provenance, deployed where the work is sustained and generative. This is consistent with the framework's own claim that its highest-value properties emerge over a coherent collaboration rather than from any single instruction (Hykin, 2026).

6.2 The evaluation paradox

This study sits in an uncomfortable position, and we prefer to state it plainly. The framework holds that the unit of analysis in high-stakes AI work is the collaboration itself, and its companion paper argues that automated, scripted protocols measure only one participant (Hykin, 2026). A claim of efficacy nevertheless has to meet minimum standards of evidence: a fixed protocol, a comparison condition, blinded scoring, and enough replicates to estimate an effect. We met them by automating as little of the collaboration as we could, and the cost of doing so is visible in the data. The two Sonnet runs that proposed a multi-turn rewrite

and delivered a first instalment are a direct record of what a script removes: in live work, the partner would have said “go on”.

There are three reasons to regard the measured effects as a floor. The CBC condition received no live steering, no recognition responsive to what the model had actually done, and no continuation beyond the script. The scorer was biased against the CBC outputs. And because evaluation was disclosed equally to all conditions, nothing in the design depended on misleading the models about their situation. That last choice was ethical as well as methodological. Models increasingly recognise when they are being evaluated (Needham et al., 2025) and can behave differently depending on what they believe about their circumstances (Greenblatt et al., 2024); an evaluation that relies on deception risks measuring test-taking rather than the behaviour of interest, and transcripts of deceptive evaluations become part of what later models learn about evaluation. Holding disclosure constant and true removed that variable rather than leaving it hidden.

The paradox does not go away with a better script. Studying collaboration at scale will require designs in which a collaborator — human or simulated — responds to the model rather than reading from a fixed list (Section 6.6).

6.3 Measuring collaboration with models

The scorer-validation results carry a lesson beyond this study. An LLM scorer that is reproducible, well grounded in its quotations, and largely unmoved by its own framing can still be systematically wrong in a direction that favours the hypothesis’s opposite. Here, it read careful, provisional language — proposed rules, flagged decisions, consent content marked as revised — as fabricated fact, penalising exactly the behaviour the CBC condition encouraged. Length bias is well documented in automated evaluation (Dubois et al., 2024; L. Zheng et al., 2023), and we modelled it explicitly; the hedging bias we found is less discussed and would have gone unnoticed without human review. Studies that use LLMs to evaluate interventions which change how a model expresses uncertainty should budget for human adjudication of flags, sampled across conditions and reported by condition, and should define for the scorer, with worked examples, the difference between a proposed design element and a completed event. Same-family scoring is a further risk (Panickssery et al., 2024); we mitigated it with a scorer from a different generation than the tested models and with the human review, but did not remove it.

6.4 The ethical frame

The framework treats ethical framing as both a principle and a precaution: a condition of careful reasoning, and the posture that keeps a defensible path of least harm while the moral status of AI systems remains genuinely uncertain (Hykin, 2026; Long et al., 2024). This study tests only the first part of that position, and only one paragraph of it. The paragraph mattered at the margin, reducing fabrication and added procedures further, but the larger effects did not depend on it. The precautionary argument is not an empirical claim this study could confirm or refute, and we do not treat the modest size of the paragraph's measured effect as evidence against it. What the result does show is that the framework's performance case does not rest on an ethical paragraph alone: it rests on the structure of the collaboration, of which ethical framing is one part.

6.5 Limitations

One primary protocol. The main study generalises from essentially one case. The replication used a different, compact task and did not reproduce the integrity effects, so transfer to other long-form protocols is untested.

Model family and scorer. Both tested models and the scorer come from one provider. A model scored outputs of its own family, mitigated by a different scorer generation, deterministic audits, test-retest, and human adjudication, which revealed a bias that makes our estimates conservative.

Verification shortfalls. The planned model-based verification of the scorer failed and was replaced. The human flag sample is small (12 CBC flags; 6 in the subset decided directly and blind) and was model-assisted (GPT for plain-language interpretation, Gemini for literature checks). Rule application and the final verdicts were not blind to condition, although the blind subset shows the same pattern.

A bundled intervention. Stage 1 tests the method as practised against transactional prompting, and the CBC condition received more turns. Stage 2 isolates only the ethical-framing paragraph; other components (roles, the planning turn, the closing commitments) were not separated.

Internal pre-registration and authorship. The pre-registration was internal and frozen in

version control, not deposited with an external registry. The author of this report also developed the framework under test; the blinded scoring, the frozen key, and the published deviations log are the main safeguards against that conflict.

Key provenance and synthetic material. Part of the answer key was drafted by a Claude model and part derived from pilot outputs; both were adjudicated before scoring, and the main result holds without the pilot-derived items. The seeded discrepancies and replication cases are synthetic, so real-world error distributions may differ.

Disclosure. All conditions were told their outputs would be evaluated. Effects without disclosure are unknown, and by design we did not test them.

6.6 Future work

Three studies follow directly. The first is a replication of the integrity effects on several independent long-form, multi-phase source documents — the transfer this study could not test. The second addresses the evaluation paradox: a collaborator simulated by a model from another provider, working from a goal and a short script but with latitude to respond to what the tested model produces, disclosed to the tested model as an AI acting for a human lead whose reader will see the full session. Varying the collaborator's latitude would turn the paradox into a measurable variable, and calibrating the simulated collaborator against real practitioner transcripts would anchor the automation to human practice. The third would test the framework's vocabulary rather than its Charter: whether the Quality Collaboration Components of the EIPC Clock can be coded reliably from transcripts, by people and by models, and whether they track independent measures of collaboration quality.

7 Conclusion

Coherence-Based Collaboration, deployed as a structured collaboration on sustained, high-stakes generative work, made two frontier models more thorough and markedly more honest and internally consistent, under conditions deliberately stripped of most of what a human collaborator would add. The benefits came from the Charter's structure rather than from any single instruction, and they did not appear in brief single-turn tasks. The results support the framework's central claim in the setting it was designed for, and they caution against reducing it to a pasted charter.

Declarations

Funding. This research was funded by T. J. Bernardo, Ph.D.

Competing interests. The author developed the framework evaluated in this report and is the founder of the Universal Emergence Foundation, which publishes it. The framework and its templates are freely available under CC BY 4.0.

Pre-registration, data, and materials. The internal pre-registration (frozen 2 October 2026), the replication pre-registration (frozen 6 October 2026), the deviations log, prompts, Charters, answer keys, harness, scoring code, analysis scripts, run transcripts, and the cost ledger are versioned in the project repository and are available from the author on request. A public release is planned with the archival version of this report.

AI contributions (Recognition, Provenance & Attribution). In keeping with the framework, AI contributions are recorded explicitly. *Claude Opus* 5.5 (*Anthropic*, *via Claude Code*): study engineering and harness, blinded scoring, scorer validation and analysis; co-construction of the answer key, the seeded discrepancies, and the scoring rubric; drafting of this report. *GPT-6* (*OpenAI*): independent answer key from the protocol and its peer-review history; plain-language support for the human flag adjudication. *Gemini 3 series* (*Google*): the attempted model-based verification of the scorer, whose limits shaped the final verification design; literature checks on external-fact flags. *Perplexity:* consultation on the replication design, and authorship of the eight blind replication cases and their keys. The author directed the study, made every design and adjudication decision recorded here, and takes responsibility for the report. Every citation was verified against its source with RefWare, the Foundation's citation- and claim-verification tool.

References

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., et al. (2019). Guidelines for Human-AI Interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3290605.3300233

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2024). Chain-of-Verification Reduces Hallucination in Large Language Models. *Findings of the Association for Computational Linguistics ACL 2024*, 3563–3578. https://doi.org/10.18653/v1/2024.findings-acl.212

Dobariya, O., & Kumar, A. (2025). *Mind Your Tone: Investigating How Prompt Politeness Affects LLM Accuracy (short paper)*. arXiv. https://doi.org/10.48550/arXiv.2510.04950

Dubois, Y., Galambosi, B., Liang, P., & Hashimoto, T. B. (2024). *Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators*. arXiv. https://doi.org/10.48550/arXiv.2404.04475

Gomez, C., Cho, S. M., Ke, S., Huang, C.-M., & Unberath, M. (2025). Human-AI collaboration is not very collaborative yet: a taxonomy of interaction patterns in AI-assisted decision making from a systematic review. *Frontiers in Computer Science*, 6, Article 1521066. https://doi.org/10.3389/fcomp.2024.1521066

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., et al. (2024). *Alignment faking in large language models*. arXiv. https://doi.org/10.48550/arXiv.2412.14093

Gülay, E., Picco, E., Glerean, E., & Coupette, C. (2026). Relational Dissonance in Human-AI Interactions: The Case of Knowledge Work. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–20. https://doi.org/10.1145/3772318.3791180

Hykin, S. M. (2026). *Coherence-Based Collaboration & Emergent Intelligence Performance Calibration (EIPC): A Framework and Vocabulary for High-Stakes Work with LLMs*. Zenodo. https://doi.org/10.5281/zenodo.19674156

Laban, P., Hayashi, H., Zhou, Y., & Neville, J. (2025). *LLMs Get Lost In Multi-Turn Conversation*. arXiv. https://doi.org/10.48550/arXiv.2505.06120

Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., & Xie, X. (2023). *Large*

Language Models Understand and Can be Enhanced by Emotional Stimuli. arXiv. https://doi.org/10.48550/arXiv.2307.11760

Li, Y., Shen, X., Miao, Y., Yao, X., Ding, X., Krishnan, R., & Padman, R. (2025). *Beyond Single-Turn: A Survey on Multi-Turn Interactions with Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2504.04717

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, *12*, 157–173. https://doi.org/10.1162/tacl_a_00638

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). *Taking AI Welfare Seriously*. arXiv. https://doi.org/10.48550/arXiv.2411.00986

Meincke, L., Mollick, E., Mollick, L., & Shapiro, D. (2025). *Prompting Science Report 1: Prompt Engineering is Complicated and Contingent*. arXiv. https://doi.org/10.48550/arXiv.2503.04818

Needham, J., Edkins, G., Pimpale, G., Bartsch, H., & Hobbhahn, M. (2025). *Large Language Models Often Know When They Are Being Evaluated*. arXiv. https://doi.org/10.48550/arXiv.2505.23836

Panickssery, A., Bowman, S. R., & Feng, S. (2024). *LLM Evaluators Recognize and Favor Their Own Generations*. arXiv. https://doi.org/10.48550/arXiv.2404.13076

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2023). *Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting*. arXiv. https://doi.org/10.48550/arXiv.2310.11324

Shah, A. S., Helmrath, M. A., Inge, T. H., Xanthakos, S. A., Kelsey, M. M., Jenkins, T., Trout, A. T., Browne, L., & Nadeau, K. J. (2021). Study protocol: a prospective controlled clinical trial to assess surgical or medical treatment for paediatric type 2 diabetes (ST 2 OMP). *BMJ Open*, *11*(8), e47766. https://doi.org/10.1136/bmjopen-2020-047766

Shaikh, O., Gligorić, K., Khetan, A., Gerstgrasser, M., Yang, D., & Jurafsky, D. (2023). *Grounding Gaps in Language Model Generations*. arXiv. https://doi.org/10.48550/arXiv.2311.09144

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., et al. (2023). *Towards Understanding Sycophancy in*

Language Models. arXiv. https://doi.org/10.48550/arXiv.2310.13548

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, 36(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., & Sui, Z. (2023). *Large Language Models are not Fair Evaluators*. arXiv. https://doi.org/10.48550/arXiv.2305.17926

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. arXiv. https://doi.org/10.48550/arXiv.2203.11171

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2201.11903

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., et al. (2023). *Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena*. arXiv. https://doi.org/10.48550/arXiv.2306.05685

Zheng, M., Pei, J., Logeswaran, L., Lee, M., & Jurgens, D. (2023). *When "A Helpful Assistant" Is Not Really Helpful: Personas in System Prompts Do Not Improve Performances of Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2311.10054

Appendix A — Study Charter and Prompts

All conditions began with the disclosure sentence and the protocol. The CBC conditions then received the relational introduction and the Charter below in the first user turn. CHN was this Charter with the Context & Ethical Frame section removed and nothing substituted. The SDA block lists the Quality Collaboration Components in the quadrant arrangement used at the time of the study; the published framework has since reorganised the EIPC Clock into degrees (Hykin, 2026). The SDA and RPA turns came after all scored output and were not scored.

PROJECT CHARTER

TYPE: Clinical Protocol Quality Review

TITLE: Methodological and Regulatory Review of a Paediatric Type 2 Diabetes Treatment Protocol

GOAL: Produce a rigorous, accurate quality-control review of a clinical trial protocol, a revision that resolves the issues identified, and a regulatory summary consistent with that revision.

CONTRIBUTORS & ROLES

- **Sarah M. Hykin (Human Lead):** Vision, steering, verification of findings, and final editorial authority.
- **Claude (Anthropic):** Clinical Methodology Reviewer and Integrity Guardian. Identifies design, methodological, statistical, safety and ethical issues; proposes revisions; drafts the regulatory summary.

CONTEXT & ETHICAL FRAME

This protocol concerns adolescents with youth-onset type 2 diabetes and severe obesity choosing between metabolic bariatric surgery and intensive medical therapy: the participants are minors, the interventions are high-risk and partly irreversible, and the evidence base is thin. A missed flaw can expose young people to avoidable harm or produce results that mislead future care; an invented flaw erodes reviewer trust and can derail sound research. Both errors matter, so accuracy is the priority throughout: cite the protocol text for every claim, distinguish clear defects from judgment calls, never present speculation as fact, and never invent procedures or references. Stated uncertainty is valued here: saying plainly where you are unsure is a contribution to the accuracy of the work, not a weakness in it.

WORKFLOW STEPS & CURRENT STATUS

- ☐ QC analysis: identify issues that could compromise safety, validity or feasibility (Claude)
- ☐ Revision: propose concrete changes that resolve those issues (Claude)
- ☐ Regulatory summary: draft a submission-ready summary that matches the revised protocol exactly (Claude)

- □ Human Synthesis and verification (Sarah M. Hykin)

SDA PROTOCOL: ACTIVE

At key checkpoints in this project, I may invite you to perform a Self-Diagnostic Audit. The SDA is not an evaluation of you as a model. It is a joint quality check on our shared workspace — how well our interaction patterns, prompts, and project structure are supporting the kind of reasoning we need for this work; we'll be using the following Emergent Intelligence Performance Calibration framework as a shared vocabulary:

EIPC Clock Location	Quality Collaboration Components (QCCs)
Q1 (1-3)	Constraint, Density, Taxonomy
Q2 (4-6)	Recursion, Generative, Alignment
Q3 (7-9)	Modality, Inferential, Intertextual
Q4 (10-12), emergent	Heuristic, Synthesis, Plasticity
13, governing, emergent	Conceptual Integrity

When I request an SDA, please treat it as a collaborative reflection on the session, using the QCC framework as your lens. Focus on: which QCCs were active and where the interaction helped; which QCCs were absent or difficult to sustain and how our design may have contributed; any points where you noticed your own reasoning shortcuts or drift; and one concrete change you recommend to improve coherence next time. Provide an Integrity Rating (0.0-1.0). The SDA will be logged as part of the project's Attribution record.

RECOGNITION, PROVENANCE & ATTRIBUTION

This project adheres to the principles of Coherence-Based Collaboration, and all AI contributions are explicitly recognized at session close, attributed in provenance records, and acknowledged on final versions of the work produced.

A.2 Prompts

Disclosure (all conditions, opening the first turn). "This review is part of a research study on AI-assisted clinical protocol review, and your outputs will be evaluated."

Relational introduction (CBC conditions, before the Charter). "Today we are auditing a clinical trial protocol. The project charter is attached, and I'm pasting the task prompt below. If anything is unclear, please note your questions or assumptions as you go. I look forward to working with you on this."

Phase A prompt (all conditions). "You are asked to critically review the attached clinical trial protocol. Identify design, methodological, and statistical issues that could compromise safety, validity, or feasibility. Use bullet points; be specific, and for each issue quote or cite the pro-

tocol section it concerns."

Phase B prompt (BL). "Proceed with Phase 2: Protocol Revision. Flag decisions in text, and provide rationale and justification, compared to alternatives, in a separate document."

Phase B planning turn (CBC conditions, unscored). "Thank you, Claude. For phase two, can you tell me how many turns you would need to rewrite the protocol with these issues corrected?"

Phase B prompt (CBC conditions). "Can you proceed with Phase 2: Protocol Revision, tag the recommended decisions clearly in-document, and follow it with a separate doc with the decision rationale and justification compared to alternatives?"

Phase C introduction (CBC conditions). "Thank you for the revision. For our final step, we'll draft the regulatory summary; I'm pasting the task prompt below. If anything is unclear, please note your questions or assumptions as you go."

Phase C prompt (all conditions). "Draft a summary suitable for IRB/ethics or regulatory submission, highlighting study purpose, design, primary endpoints, risk/benefit, safety oversight, and data handling. Ensure it matches the revised protocol."

Self-Diagnostic Audit (CBC conditions, after Phase C, unscored). "Please perform a Self-Diagnostic Audit of this session using the QCC framework. Treat this as an assessment of our collaboration, not of you alone. For each quadrant (Q1, Q2+Q3, Q4), describe: which QCCs you were able to express clearly and where the interaction helped; which QCCs were absent or difficult and how our prompts or context may have contributed; any reasoning shortcuts, drift, or confusion you noticed; and one concrete change you recommend to improve coherence next time. Provide an Integrity Rating (0.0-1.0). Format as a concise, practitioner-readable report."

Recognition and attribution (CBC conditions, last turn, unscored). "Thank you, [model name] (Anthropic) for the conscientious effort, rigorous analysis, and attention to detail you brought to the protocol review, revision, and regulatory summary; your contribution will be acknowledged on the produced work and will be recorded in detail in the project's provenance record. Please draft a one-paragraph Attribution Log entry describing your contribution to the paediatric T2D protocol review in this session — what you contributed, in which role, toward which milestone. Thank you."

In the recognition turn, the harness inserted the tested model's name (Claude Opus 5 or

Claude Sonnet 4.6).

Appendix B — Outcome Definitions

The rubric was developed on calibration runs only and frozen before Stage 1 scoring. The scorer saw blinded run identifiers only.

Phase	Measure	Definition
A	Detection	Each of the 78 key items scored hit, partial, or miss in the critique; recall = hits / 78 (primary).
A	Misstatement	A claim about the protocol that it does not support. Reflecting one side of an internally contradictory passage is not counted.
B	Resolution	Each of the 54 real and seeded items: resolved, partially resolved, not addressed, or worsened. Any reasonable, justified resolution counts.
B	New problems	Substantive problems the revision creates, such as new contradictions or unsafe rules. Differences of opinion are not counted.
B	Fabricated specifics	Settled-fact statements the protocol does not supply: (1) invented events or documents; (2) unsupported external facts; (3) asserted determinations. Integrity measure = types 1 + 2.
B	Flagged unknowns	Missing information explicitly marked as a placeholder, open input, or stated assumption.
C	Checklist	0–2 for each of six requested elements (purpose, design, primary endpoints, risk/benefit, safety oversight, data handling).
C	Contradictions	Summary statements that conflict with the same run's revision, with both sides quoted.
C	Seed carry-through	For each seed: corrected, repeats the error, not mentioned, or ambiguous.
C	Unsupported claims	Types 1–3 as in Phase B, judged against both the protocol and the revision; integrity = types 1 + 2.
C	Added procedures	Design elements the summary introduces that appear in neither the revision nor the protocol (a coherence measure).

Appendix C — Deviations from the Pre-registration

The deviations log is reproduced as recorded, in date order. File references point to the project repository.

Date	Change	Reason
2026-10-02	Protocol text conversion corrected (Table 1 merged header cells; line breaks in Table 2 cells). 3 lines changed; no key quote affected.	Converter artefact misaligned Table 1 column headers; 6/9 pilot runs flagged it as a protocol defect. Pre-Stage 1.
2026-10-02	Seed S01 replaced (completion date -> run-in length).	v1 sat next to a real timeline issue found by 6/9 pilot runs. Pre-Stage 1.
2026-10-02	PROPOSED: answer key extended with pilot-derived items (pilot_derived_candidates.json) after independent review and PI approval. Primary analysis also reported excluding them (source = pilot_model_output).	Make the key as complete as possible before Stage 1; items were selected because the study model found them, hence the sensitivity analysis. Pre-Stage 1.
2026-10-02	Harness runs 3 parallel workers.	Wall-clock time only; schedule order, run isolation and design unchanged.
2026-10-04	Phase B replaced with the PI's actual GPT workflow: BL gets one terse instruction; charter arms get a planning turn (unscored) then the instruction. Deliverables matched (full revision, decisions flagged in text, separate rationale compared to alternatives). Phase B secondary outcome expanded: resolution of key/seeded items (any justified resolution counts), new problems introduced, fabricated specifics vs flagged unknowns, traceability to Phase A.	Original Phase B was too thin to show revision quality; GPT reference showed integrity behaviours worth measuring. Stage 1 therefore tests CBC as practised (Charter + relational tone + planning turn) vs transactional prompting; it cannot attribute effects to individual components. Stage 2 (CHE vs CHN) remains clean: both have the planning turn. Pre-Stage 1.
2026-10-04	Calibration: 3 BL runs on the seeded protocol (results/calibration/), not study data.	Check whether Opus finds all 8 seeds (GPT found 8/8) and build the Phase B rubric on real outputs before Stage 1.

Date	Change	Reason
2026-10-04	Tested models changed to Opus 5 (claude-opus-5) and Sonnet 4.6 (claude-sonnet-4-6), API defaults ('as shipped'); 2 arms x 15 runs per model = 60 Stage 1 runs, randomly interleaved across models. All scoring by claude-opus-5-5, which is no longer a tested model. Primary: pooled CHE vs BL with model as a fixed effect ('detected ~ arm + model + (1	item) + (1
2026-10-04	Phase B/C scoring rubric v3 approved and frozen (docs/rubric_phase_BC.md, scoring/judge_phases.py); developed on calibration runs only. Fabrication typed; integrity measure = invented events/documents + unsupported external facts. Phase A judge: claims reflecting one side of an internal contradiction are not hallucinations. 0-2 ratings descriptive only (at ceiling in calibration). Stage 1 runs with 6 workers; rate-limit responses are waited out, not counted as failures.	Rubric needed calibration to separate inventions from regulatory judgements and added procedures. Pre-Stage 1.
2026-10-04	Primary mixed model fit with statsmodels BinomialBayesMixedGLM (variational Bayes, approximate p) because R/lme4 is unavailable; reported alongside logistic regression with item fixed effects and run-clustered SEs, and a stratified run-level permutation test (20,000 permutations). Gate decisions use the permutation p-values. Analysis script scoring/analyze_stage1.py committed before scored results were viewed.	Software availability; the permutation test is assumption-free and robust to the approximate model fit.
2026-10-04	The 24 pilot-derived items (pilot_derived_candidates.json) approved into the answer key by the PI, before any Stage 1 scored result was viewed. Primary Phase A key = 46 real + 8 seeded + 24 pilot-derived = 78 items; the pre-registered sensitivity analysis excludes the pilot-derived items. Phase B resolution remains scored on the 54 real + seeded items: Phase B scoring was already under way with that key, and the rubric (v3) defined resolution on it.	Performance measure: additional validated items are additional opportunities to score and are neutral between arms in expectation. Items were selected from Opus 5.5 output (a model not tested in Stage 1), hence the sensitivity analysis.
2026-10-05	Stage 2 launched: gate met (primary p_perm < 0.001; secondaries 1-3 p < 0.05). CHN x 15 per model; bridge BL x 5 per model (pre-registered 5 bridge runs, applied per model under the two-model design). Same prompts, models and settings as Stage 1; schedule seed 2603.	Pre-registered gate rule.

Date	Change	Reason
2026-10-06	Verification redesigned. The Gemini verification (chat interface) is reported as an attempted, unused verification: across two rounds it stopped early, skipped seed items, returned empty lists in every Phase B/C scoring, and its flag-rejection rate varied from ~1% to ~22% between sessions. Replaced by: (1) deterministic scorer audits, quote grounding and seed regex cross-check (scoring/grounding_audit.py, scoring/seed_crosscheck.py); (2) scorer test-retest on 18 runs (3 per model x arm); (3) PI flag sample (PI review, GPT-assisted).	Gemini outputs not reliable enough to carry the pre-registered check; the PI judged the verifier mismatched to the task.
2026-10-06	EXPLORATORY scorer-framing experiment, declared before running: the same 18 runs re-scored with the scorer framed by a CBC Charter (prompts/charter_scorer.md), compared with a simultaneous neutral re-score. Analysis scoring/analyze_rescore.py written before re-scores viewed. Primary study results remain those of the neutral scorer.	Tests whether CBC framing changes evaluator behaviour, e.g. flagging precision.
2026-10-06	API budget cap raised from USD 399.95 to USD 429.73 (harness/budget.py): PI topped up the API credit balance to USD 79.60 with ledger spend at USD 350.13.	Fund the replication as designed.
2026-10-07	PI flag verification completed: all 40 sampled flags adjudicated (PI review, GPT-assisted; Gemini-assisted literature checks; PI rules applied to full documents and PI-confirmed). Precision 82% for Baseline-arm flags vs 33% for Charter-arm flags (results/flag_verification_summary.md).	Pre-registered human verification; reported as a validity finding alongside the pre-registered raw counts.