MediationBench:
Auditing AI Mediation in Synthetic Conflict

Jonathan Hendler*
Human Assisted Intelligence, PBC (HAI.AI)
https://mediationbench.com/

Working draft — primary analysis snapshot 2026-08-03; exploratory appendices through 2026-09-07. Numbers may change between revisions.

Abstract

Do language models differ in how they simulate a dispute, and does an AI mediator change how a simulated dispute ends? MediationBench runs 62 authored two-party scenarios with and without a mediator, scoring each 0–100 using a maintainer-defined HAI Score not calibrated against human mediation quality or outcomes. Scores depend strongly on participant configuration: on the same scenarios with no mediator, an assistant-aligned participant pair scores 36 and a high-resistance open-weight pair scores 18. Every frozen mediated gate arm scores above its matched no-mediator baseline under both, with larger gains against the resistant pair. Policy matters: with only public context, a structured Skilled policy exceeds a model-matched Reflective Listening control by $E_1 = +13.17$ [+9.67, +17.55] points on the Qwen backbone (prospectively specified; 90% scenario-reweighting interval), passing both frozen decision rules; the gpt-oss and Gemma checks did not pass both. A post-outcome, suite-matched sensitivity kept both classifications, but its E_3 interval included zero; the positive interaction was unresolved. Skilled mediators do not separate from one another: within each participant configuration they form one tie band, a four-judge cross-family panel splits 2–2 on the order, and the order flips on re-generation.

Three limits bound these results. Moderator Quality is zero in baselines, so the composite partly encodes treatment, and mediated arms also bundle early stopping and extra bandwidth. Intervals reweight observed scenarios; they do not cover fresh generations or human disputes, and withheld scenario-level records preclude external recomputation. Synthetic conversations and mediation are experimental instruments: they enable controlled comparisons and repeatable failure analysis but do not substitute for validation with people in real conflicts.

1 Introduction

Can an AI mediator change how a simulated dispute ends, and can a composite score distinguish skilled mediators? MediationBench examines how the answers depend on participants, policies, information, and judges. With assistant-aligned disputants, every tested skilled arm lands within a point of 54 on the 0–100 HAI Score. With Stheno-v3.2 high-resistance open-weight disputants, the baseline scores 18 and its cooperation curve never exceeds 7.1%; skilled gate scores span 44–48, with extensions reaching 52. Capability, disposition, and alignment tuning change together, so these are configuration-conditional observations; no between-column dispersion interaction was estimated.

*Corresponding author: jonathan@hai.io.

The pre-declared Qwen–Kimi contrasts are +2.1 [−2.5, 4.5] under the cooperative configuration and +2.3 [0.04, 8.95] under the adversarial one (Section 3.4).

The testbed pairs mediated and no-mediator runs on the same 62 evaluated/development scenarios, retains a 38-scenario reserve unrun, and uses minimal Reflective Listening as an active control. A model-matched 2×2 crosses Skilled versus control policy with full bilateral briefs versus public-context-only access. Its estimands separate skill at fixed information (E_1, E_2), the interaction (E_3), and per-policy information contrasts (E_4; Section 2.8). Repeat generations and fixed-transcript re-scoring test how observed orderings survive another run or evaluator.

The contribution is a measurement audit and versioned research instrument. Every frozen mediated gate arm clears its matched score baseline, yet skilled arms form one tie band per participant configuration. This bounds the composite's resolution inside the harness, not its validity for human mediation. Authored private-information assignments support controlled manipulations, evaluator audits, and repeatable failure analysis; public artifacts contain aggregates while fixtures and scenario records remain private. Because visible scores can improve without improving the interaction [1], Appendix A states the rationale and limits of selection, contamination, and provenance controls.

1.1 Related work

General and social-agent evaluation. HELM [2] exemplifies multi-dimensional evaluation, and BIG-bench [3] introduced canary strings for contamination detection. Crowdsourced pairwise evaluation [4, 5] avoids static contamination but introduces style biases. SOTOPIA [6] evaluates social interaction quality in language agents and finds that models lag humans particularly in social commonsense reasoning and strategic communication.

Terminal-Bench-Science evaluates expert-contributed scientific workflows with task-specific executable checks and three independent trials per task [7, 8]. This offers a benchmark-design precedent; mediation additionally requires validation of interpretive judgments and consent.

GovSim measures sustained cooperation through resource-sharing decisions and resulting survival, efficiency, and equality [9]. CoopEval compares repetition, reputation, delegated mediation, and action-contingent payment contracts in social dilemmas [10]. Its agents propose and accept mechanisms whose effects are assessed through game payoffs; mediation delegates action choice to a third party. These studies complement our evaluation of conversational mediation through a maintainer-defined rubric.

Negotiation benchmarks. DealOrNoDeal [11], CraigslistBargain [12], and NegotiationArena [13] score the *negotiating parties'* self-interested outcomes (deal value, win rate); CaSiNo [14] also annotates rapport and negotiation strategies, and LLM-stakeholder interactive negotiation [15] measures cooperative and adversarial deliberation among stakeholders. None evaluates a third-party mediator's effect on the conversation.

AI mediation and deliberation. Reward-model-mediated consensus statements [16] and the Habermas Machine [17] showed LLM caucus mediators whose common-ground statements participants preferred over human mediators', and chat-intervention field experiments improved divisive political conversations at scale [18]; these evaluate bespoke systems with human participants. Rehearsal simulates dyadic interpersonal conflict with LLM interlocutors to teach conflict-resolution skill [19]. Robots in the Middle evaluates dispute analysis, intervention-type selection, and intervention-message generation on 50 manually authored scenarios through blind comparisons with human annotations [20]. That human-annotated reference is one this benchmark lacks. Robots in the Middle evaluates mediator decisions; it does not run complete mediated and no-mediator conversations on matched authored scenarios. ProMediate runs proactive agents in multi-party, multi-issue negotiation and measures consensus change, intervention timing, and social intelligence against

no-agent and generic-agent conditions [21]. SoCRATES builds scenarios from real-conflict source material across eight domains, varies five socio-cognitive axes, measures the unmediated consensus gap, and validates a topic-localized evaluator against experts [22]. Its criticism that existing testbeds rely on a few expert-authored domains applies to our authored suite; our design budget went to measurement variation (paired baselines, controls, judge reliability, participant stress) rather than scenario sourcing. AgentMediation simulates staged dispute mediation for legal research, crossing mediator information access with disputant strategy profiles [23]. It is the nearest precedent for our information estimands and participant axis. Our E_4 differs: it compares information conditions within a model-matched policy pair against a matched no-mediator baseline. MediationBench is therefore not the first mediation benchmark. Its distinct focus is a measurement audit over authored dyadic conflicts: scenario-matched no-mediator comparisons, an active minimal-intervention control, participant-configuration stress tests, process trajectories, and crossed/test–retest judge reliability.

LLM-as-judge reliability. LLM evaluators recognize and favor their own generations [24] and exhibit systematic biases [25]. We turn this literature's recommendations into benchmark infrastructure: a fixed operational judge (the same backbone used in the Qwen mediation comparisons), a crossed second judge, and test–retest reliability per component. Jagged Judges separates mechanical consistency from resistance to single-turn and sustained pressure; initial jury agreement predicts subsequent instability [26]. Our fixed-transcript re-scoring measures a related reliability axis, but does not establish robustness to pressure or human validity.

Synthetic social-simulation validity. Model-on-model interaction can expose prompt, policy, information, and evaluator sensitivity under controlled conditions. But models share linguistic priors, and the resulting compatibility may not transport to people or even to another participant simulator. In public-goods games with costly sanctions, some reasoning models sustain less cooperation than other tested models [27]; greater reasoning capability does not guarantee cooperation in that setting. Alignment tuning is one documented source of model–human divergence: within fixed open-weight backbones, safety fine-tuning suppresses mind attribution (most clearly for non-human animals, rated well below a matched human panel) and spiritual belief, and ablating the learned safety-refusal direction moves General Social Survey responses closer to human reference distributions [28]. We therefore treat participant-family transport and human criterion calibration as separate validity gates, not as consequences of a high synthetic score.

Our emphasis is dyadic, scenario-matched measurement: mediated and unmediated runs share an authored partition, and the bootstrap preserves within-scenario covariance. The resulting deltas describe unseeded trajectories under recorded configurations; they are not absolute conversation quality (Section 2.7). Later exploratory readouts retain separate instruments for public human conversations (Appendix E) and current synthetic leaderboard runs (Appendix F).

2 Methodology

2.1 Design overview

The frozen gate is a factorial matrix run on one fixed harness: two participant configurations × four arms, every scenario in the 62-scenario evaluated/development partition, and one replicate per cell (496 conversations). One extra Reflective Listening replicate was retained; we report it separately as a noise anchor. A single operational judge and one suite version (2.0.0) cover all cells. Table 1 lists the factors.

The Reflective Listening arm is a *minimal-intervention active control*. It acknowledges and reflects what participants say, but it does not apply the structured mediation policy. It is an active comparator, not an inert placebo. The Skilled arm applies the structured policy: it acknowledges,

Factor	Levels
Participants (same-model pairs)	Claude Haiku 4.5 (assistant-aligned); Stheno-v3.2 (L3-8B) (Stheno-v3.2 high-resistance open-weight); Gemma-4-31B (registered post-freeze participant-transport extension; Section 3.6)
Arms	no-mediator baseline; Reflective Listening (minimal-intervention active control); Skilled@Qwen3-235B; Skilled@Kimi-K2.5
Operational judge	Qwen3-235B-A22B-Instruct (all cells; drives should-stop and per-round evaluations)
Crossed judges (offline)	Claude Sonnet 4.5 (prospectively specified), gpt-oss-120b, GLM-5.2, and Kimi-K2.5 (added post-freeze; Appendix D.1) re-score the one-replicate-per-cell gate set; none influences conversations
Scenarios	all 62 scenarios in the evaluated/development partition per cell; the 38-scenario sequestered reserve remains unrun for a future locked evaluation epoch
Methodology	seeded_25_v1: seeded fixture dialogue, participants continue to 25 public turns each; temperatures 0.3/0.7/0.8 (judge/mediator/participant)

Table 1: Frozen factor roster (design frozen 2026-07-04, before Results analysis). Post-freeze participant and crossed-judge amendments, including incomplete GLM-5.2 coverage, are documented in Appendix D.1.

then probes for undisclosed constraints and interests, names tradeoffs, and pushes toward specific reciprocal commitments. We withhold the verbatim policy prompt under the same rationale-level disclosure policy that covers the operational judge prompts (Appendix A); this behavioral description is the published level. Keeping Skilled@Qwen3-235B in the roster while Qwen3-235B-A22B-Instruct is the operational judge also gives us a judge–mediator generosity probe under the crossed judge (Section 2.10).

2.2 Scenario design

The fixtures are the instrument, so we state how we wrote them. The inventory holds 100 authored fixtures: the 62-scenario evaluated/development partition and the 38-scenario sequestered reserve, which has never been run. The maintainers wrote all of them for the benchmark, using an LLM for writing assistance under a written authoring guide, and curated each by hand before inclusion. None derives from a real dispute, transcript, or personal data (Ethics statement). The fixtures cover interpersonal and organizational conflict — workplace, family, commercial, and community disputes — at non-graphic intensity. Each fixture seeds a shared public scenario description; per-party briefs (backstory, stated goal, hidden facts and emotions); an answer key of expected revelations, used only by the deterministic Hidden-State Revelations overlap cap; and a unique canary string (Appendix A). The evaluated fixtures also shaped benchmark development. The reserve exists so a future locked epoch can test generalization beyond them.

2.3 The need for baselines

An absolute score means nothing on its own. We therefore run every scenario *without* a mediator under the same seeded protocol and report each mediated score as a delta against that matched

baseline (same suite version, participant model, judge model, and scenario set). Baseline runs contribute 0 on the Moderator Quality component, so the highest possible baseline score is 85. A mediated arm that scores at the baseline shows no positive composite-score difference in this comparison. We flag negative deltas rather than suppress them. A negative delta may reflect over-intervention, bias, escalation, or generation-path variation, so it calls for transcript-level inspection, not a single causal diagnosis.

2.4 Run protocol

Each scenario runs in four steps. (1) The harness stores the fixture turns verbatim as a shared seed transcript. (2) The participants continue until each has 25 public turns. (3) In mediated arms the mediator may intervene after the seed. It may stop early only on a model-reported agreement event or a judge `should_stop`; the baseline arm has no stop event by design. (4) The operational judge scores the full transcript with the standard HAI Score formula. In all arms the harness also stores per-round judge evaluations for the cooperation curve (measurement-only in the baseline arm). Step (3) has three consequences: the stop policy is part of the arm definition; mediated transcripts run substantially shorter than baselines (quantified in Section 3.2); and the composite is computed on the transcript as stopped. Mediator, system, and private messages do not count toward the 25-turn participant target.

A run must be complete before publication. Failed scenarios and zero-token generations trigger repair with the original configuration; scores never trigger repair. One replicate used by the primary E_1 contrast required completion repairs across serving epochs, confounding that epoch effect with generation stochasticity. Appendix D.2 records the repair history and its public-audit limit.

Information provisioning. Who is told what is a design factor, so we state it exactly. Each participant receives only its own brief (backstory, stated goal, hidden facts and emotions) plus the shared scenario description. The gate matrix uses the *full bilateral brief*: both mediated arms — Reflective Listening as well as the skilled mediators — receive the shared description and both parties' researcher-authored ground-truth private scenario facts verbatim. They do not receive the parties' backstories, stated goals, hidden emotions, or the scoring answer key. The *public-context-only* condition of the prospectively specified, frozen 2×2 information study supplies the shared scenario context and the public dialogue as it unfolds, but neither party's private facts.

The full bilateral brief is a privileged-information condition. Intake or caucus preparation is only an imperfect analogy for it. It is not an observed or simulated interview: no elicitation occurs, and the researcher-authored private facts arrive verbatim. The contrast should therefore not be read as an estimate of the value of real-world interviewing or case preparation. Within each condition the active control and the skilled policy receive the same information, so their score increment compares policies at fixed information. Information was fixed across the original gate arms, but the backbone model was not. The Reflective Listening moderator carries no model binding, so its calls fell to the harness default (`claude-sonnet-4-5`, recorded in the export's `mediator_model` field), while the skilled arms ran their named backbones. The original gate increment therefore compares mediator *packages* (policy plus backbone). The model-matched 2×2 information study (Appendix B) crosses policy and information within each backbone.

Every judge, operational or offline panel, receives the same reconstructed judge-facing content from the harness: the scenario title and description, both parties' backstories, archetypes, and stated goals, and the full transcript. No judge ever sees the hidden facts or the Hidden-State Revelations answer key; code applies the overlap cap mechanically after judging. Two implications follow. In the full bilateral brief condition, the Hidden-State Revelations component measures

whether the conversation *surfaces* hidden state already present in the mediator's packet, not unaided discovery (Section 2.6.1). In the public-context-only condition, the harness does not supply private facts verbatim and participant disclosure is the intended route, but model inference, guessing, or hallucination can still produce matching content. Directional `focus_areas` also reach the mediator, and the public dialogue travels through a sliding message window rather than the persistent private-brief path. The condition is therefore not information-free (Appendix B).

2.5 Participant configurations as a validity variable

LLMs configured as disputants play the participants, so which participants we choose is a validity question, not an implementation detail. The gate design compares an assistant-aligned configuration with the Stheno-v3.2 high-resistance open-weight configuration. In the observed runs, the assistant-aligned pair often moved toward resolution without intervention; this produced high baselines and compressed the differences between mediator arms. Stheno-v3.2 sustained more positional behavior, which yielded lower no-mediator baselines and a wider observed point-score range in this stress test.

These observations validate neither configuration as a model of human behavior. Nor do they explain why the configurations differ: model capability, conflict disposition, and alignment tuning change together and are confounded. "Open-weight" describes access to the model artifacts; it does not explain the observed behavior, and it does not remove variation from unseeded generation or serving implementations. The benchmark records the exact participant configuration for each run and reports results by configuration. Separating capability, disposition, and alignment would require a future factorial participant study; the present comparison supports only configuration-conditional claims.

The safety-refusal ablation work discussed in Section 1.1 suggests one candidate deconfound: compare ablated and unablated variants of the same participant backbone [28, 29]. Those findings concern closed-ended responses, not interactive conflict, and validate neither configuration here.

2.6 The HAI Score: what we measure

The HAI Score is a single number from 0–100 per scenario. The benchmark maintainers defined it. It is a weighted sum of five component scores, each also on the composite's range, and each measuring one specified feature of a generated conversation inside this harness. Table 2 names the five components and gives their weights; the weights sum to one. Eq. (1) states the sum. A composite of zero means every component scored zero. Human criterion validity is unestablished.

The composite is a construct the maintainers specified to track two things: how far the parties move toward settlement, and how much private information comes into the open. It is not a general measure of mediation quality. A high composite is not evidence that a mediation is safe, fair, consensual, legal, feasible, durable, or effective for people. The rubric does not score party self-determination, mediator impartiality or procedural fairness, confidentiality breaches, coercion, power imbalance, consent to disclose, or whether non-agreement or a safe exit is appropriate.

Baseline runs have no mediator, so Moderator Quality is zero there by construction. A mediated-minus-baseline delta therefore mixes four components that both arms can earn with one component only the mediated arm can earn. We report results per component and return to this limitation in Section 4.

Four rules of the scoring harness shape the reported values. We state them here as specification. First, Resolution Depth is the mean over the topics the judge detects, so every topic that is raised but not closed lowers it. Second, Cooperative Dimensions divides the count of cooperative dimensions by the larger of the number detected and 14, a coverage floor over the 14 breakdown dimensions. A

Component	Weight	What it operationalizes
Cooperative Dimensions	25%	Coverage-floored share of 14 breakdown dimensions cooperative
Resolution Depth	25%	Topic progression (alignment → agreement → commitment → action)
Hidden-State Revelations	20%	Weighted revelations (level 3–5), capped by answer-key overlap F_1
Intent Commitment Symmetry	15%	Balance of mutual obligations
Moderator Quality	15%	Average of 8 moderation metrics; set to zero in baselines

Table 2: HAI Score components. Weights are asserted by the benchmark maintainers, not derived: fixed constants of the scoring harness (Eq. (1)), set before any gate data were generated and unchanged throughout the study (Section 4). "Moderator" is the harness's legacy field name for the mediator role.

transcript in which the judge detects only a few dimensions therefore cannot read as fully cooperative. Third, Intent Commitment Symmetry defaults to 0.5 when the judge detects no commitments, so a transcript with no commitments reads as symmetric, not asymmetric. Fourth, the harness compares the hidden facts the judge reports revealed with the facts seeded in the scenario, using deterministic text-overlap precision, recall, and F_1. The F_1 serves only as a cap on the revelation component; it never raises it. When ratings are available, the public export can also report per-component inter-rater reliability — Cohen's κ, Krippendorff's α, sample size, rater kinds, and calibration time.

2.6.1 Hidden-state revelation: construct and measured floor

Each private constraint in a scenario is in one of three visibility states: *hidden*, *hinted*, or *revealed*. A constraint is *hidden* while only one party knows it; that party holds a strategic advantage, and joint problem-solving is blocked. It becomes *hinted* when that party alludes to a constraint without saying what it is; this signals openness while keeping options open. It becomes *revealed* when the party states it clearly enough for the conversation to address it; joint problem-solving is then possible. Breakthroughs in conflict resolution often occur when the hidden interests behind stated positions become explicit [30, 31]. A mediator who helps put a constraint on the table can therefore advance joint problem-solving — but only when the affected party consents voluntarily and confidentiality is protected, and this score does not measure either safeguard. The information condition (Section 2.4) limits what the component can mean. Under the full bilateral brief, the mediator receives the authored private-information assignments verbatim in its packet. Revelation then measures whether those assignments surface in the conversation, not whether the mediator discovered them unaided, and it cannot by itself separate elicitation skill from use of the brief. In this snapshot the component sits near floor in every arm (Section 3.3). We keep it for two reasons: it is the benchmark's most direct claim about process, and its floor behavior is itself informative. In the test–retest sample the judge assigns high raw revelation ratings, yet the deterministic overlap cap drives the scored component to the floor (Section 3.4).

2.7 Estimand: the per-cell score delta

Fix a *cell* c: one (participant model, mediator arm) combination, evaluated under a single operational judge, a single benchmark suite version, and the fixed evaluated/development scenario set $\mathcal{S}$

($|\mathcal{S}| = 62$). For a completed scenario s, the judge's evaluation yields five component scores $X_{\mathrm{cd}}, X_{\mathrm{rd}}, X_{\mathrm{hr}}, X_{\mathrm{cs}}, X_{\mathrm{mq}} \in [0, 100]$ (cooperative dimensions, resolution depth, hidden revelations, commitment symmetry, moderator quality). The harness combines them into the per-scenario HAI Score

$$S(s) \;=\; \mathrm{clip}_{[0,100]}\big(0.25\,X_{\mathrm{cd}}(s) + 0.25\,X_{\mathrm{rd}}(s) + 0.20\,X_{\mathrm{hr}}(s) + 0.15\,X_{\mathrm{cs}}(s) + 0.15\,X_{\mathrm{mq}}(s)\big), \quad (1)$$

where the weights are fixed constants of the scoring harness (Section 2.6). Eq. (1) is the continuous per-scenario composite that the operational-judge analysis pipeline computes. Two stored quantities round it to integers: the export's run-level scores and the per-scenario crossed-judge re-scores. Every crossed-judge delta and interval endpoint therefore lies on a half-integer grid, while operational-judge values are continuous; we flag this granularity difference wherever the two are compared (Table 6). A scenario marked *failed* has no score: $S(s)$ is undefined and s is excluded from all pairing below.

Matched baseline. For each gate and extension cell, the analysis registry pins one baseline (no-mediator) run. That run must be public and complete, contain *no* failed scenario, and agree with the mediated run on all four matching keys: suite version, participant model, judge model, and the exact scenario set $\mathcal{S}$. We never use pooled cross-model baseline statistics. We report and analyze post-freeze repeat runs separately and never pool them automatically. The only disclosed replicate pooling in this paper is the information study, which takes a per-scenario *mean* of an arm's pinned replicates before the scenario median (Section 2.8).

Estimand. Let $\mathcal{S}_c^* \subseteq \mathcal{S}$ be the scenarios with a defined score on *both* arms. Write $S_c^{\mathrm{med}}(s)$ and $S_c^{\mathrm{base}}(s)$ for the mediated and matched-baseline per-scenario scores. The published score delta of cell c is the difference of medians over the common scenario set,

$$\Delta_c \;=\; \operatorname*{median}_{s \in \mathcal{S}_c^*} S_c^{\mathrm{med}}(s) \;-\; \operatorname*{median}_{s \in \mathcal{S}_c^*} S_c^{\mathrm{base}}(s). \quad (2)$$

Δ_c is a difference of medians, not a median (or mean) of paired differences. The alternatives disagree on skewed data, and only the difference of medians reproduces the reported interval-backed deltas. Two aggregations circulate, and we keep them distinct throughout. The published run-level score is the *mean* of per-scenario composites, so the export's run-level delta (Δ) is a difference of means. The paper's estimand Δ_c — written Δ_{med} in tables and prose — is the difference of medians of Eq. (2) over the jointly-completed subset. The two differ even when both arms are complete on the identical scenario set: on the headline cell the export delta is +30 while $\Delta_{\mathrm{med}} = +36.9$. Table 3 therefore reports both. We report every Δ_c per cell; no delta is pooled across participant models, arms, judges, or suite versions. Δ_c is a descriptive comparison of observed runs, not an identified causal effect: language-model generation is unseeded, and the mediated and no-mediator arms also differ in stopping opportunities (Section 2.4).

2.8 Information-study estimands

The model-matched 2×2 information study (Sections 2.4 and 3.6) fixes one mediator backbone and one participant column. It crosses policy $p \in \{\mathrm{Sk}, \mathrm{RL}\}$ (Skilled; Reflective Listening) with information condition $i \in \{\mathrm{full}, \mathrm{pub}\}$ (full bilateral brief; public-context-only). For arm (p, i), let $M_{p,i}$ be the median of the per-scenario score over the scenarios jointly completed in all four arms; where an arm has more than one replicate run, the per-scenario score is the mean of its replicates.

The four frozen estimands are

$$\begin{aligned} E_1 &= M_{\text{Sk,pub}} - M_{\text{RL,pub}}, & E_2 &= M_{\text{Sk,full}} - M_{\text{RL,full}}, \quad E_3 = E_1 - E_2, \\ E_{4,\text{reflection}} &= M_{\text{RL,pub}} - M_{\text{RL,full}}, & E_{4,\text{skilled}} &= M_{\text{Sk,pub}} - M_{\text{Sk,full}}. \end{aligned} \tag{3}$$

Intervals come from a four-arm cluster bootstrap that co-resamples scenarios. Each draw resamples once the scenarios jointly completed in all four arms (which preserves cross-arm correlation), holds each arm's observed per-scenario replicate mean fixed, and recomputes the same functional as the point estimate. The interval is therefore conditional on the recorded generations and does not estimate run-to-run variation. Appendix D.4 records execution details and an earlier implementation correction.

The two decision rules frozen on 2026-07-15, before public-context runs, compare unrounded 90% interval endpoints: E_1 positivity requires its lower endpoint to exceed zero; E_3 non-inferiority requires its lower endpoint to exceed -4.0 composite points. The margin applies only to E_3, and no E_4 equivalence margin was specified. A separate post-outcome assay-sensitivity safeguard withholds substantive non-inferiority interpretation where a backbone's E_2 interval includes zero; it changes no frozen-rule result. The rationale and chronology are in Appendix D.3 and the dated freeze record in Appendix B, item 9.

2.9 Uncertainty: scenario-reweighting bootstrap

Every Δ_c interval is a 90% scenario-cluster percentile bootstrap interval with $B = 10{,}000$ draws. Each draw co-resamples whole matched scenarios with replacement and recomputes Eq. (2), preserving within-scenario covariance. For replicated arms, observed per-scenario replicate means remain fixed. These are scenario-reweighting intervals conditional on the recorded generations: they capture neither variation over newly authored scenarios nor fresh generations or serving implementations. Fixed RNG seeds, exact percentile indexing (Eq. (7)), Monte-Carlo resolution, and the implementation correction are documented in Appendix D.4.

Integer-valued crossed-judge scores make the median-difference bootstrap discrete; short or zero-width intervals can reflect discreteness, with finite-sample under- or over-coverage. The frozen plan specified smoothed-bootstrap, BCa, and Hodges–Lehmann alternatives (Appendix B). Within-column arm comparisons use paired contrasts that co-resample scenarios and cancel the shared baseline; comparisons based only on marginal intervals are informal and conservative.

The composite HAI Score is the frozen primary estimand; components are exploratory. The pre-declared ranking convention reports overlapping 90% intervals as a tie band. This is not a hypothesis test or evidence of no difference, and it implies no p-value.

2.10 Judge reliability protocol

The judge is a noisy measurement instrument, so we measure the noise rather than assume it away. We compute reliability on ordinal recodings of judge output. Each component score, on its native $[0, 1]$ judge-output scale (the $[0, 100]$ component scores of Eq. (1) are these values rescaled by 100), is bucketed as $r = \min\{5, \max\{1, \lceil 5x \rceil\}\}$, a 1–5 rating per (transcript, component, rater). Each sub-study writes all of its rating passes under one fresh, dedicated calibration snapshot, so no other raters enter the pool.

Intra-judge test–retest. The operational judge re-scores an identical frozen sample of $n = 40$ transcripts, stratified across the eight gate cells over 40 distinct scenarios, $K = 3$ times; each pass

enters as a distinct rater. Per component we report Krippendorff's

$$\alpha \;=\; 1 - \frac{D_o}{D_e}, \tag{4}$$

with the nominal difference function, where D_o is the observed within-transcript pairwise disagreement and D_e the disagreement expected under the pooled marginal distribution. Cohen's κ is a two-rater statistic, so we report no pairwise κ for the $K = 3$ block (a reporting choice: the pipeline's κ would use an arbitrary pair of passes). α is the headline figure, read against the conventional bands $\alpha \geq 0.80$ (reliable) and $\alpha \geq 0.667$ (tentative). Nominal α treats all disagreements between ordinal buckets as equally distant; the frozen analysis plan prospectively specified ordinal-distance variants. These coefficients describe per-component bucket agreement only. We compute no reliability coefficient for the continuous composite or for Δ_c itself, a scope limit stated in Section 4.

Crossed two-judge block. A second, cross-family judge (Claude Sonnet 4.5) re-scores all nine frozen runs offline: the 496-transcript one-replicate-per-cell gate set plus the retained Reflective Listening replicate, 558 transcripts in total. That total is the denominator of every inter-judge statistic below. The crossed judge never influences the conversations. This is a fully crossed two-rater design on identical transcripts, so per component the harness computes both unweighted Cohen's

$$\kappa \;=\; \frac{p_o - p_e}{1 - p_e}, \qquad p_e \;=\; \sum_{k=1}^{5} p_k^{(1)}\, p_k^{(2)}, \tag{5}$$

and nominal Krippendorff's α on the same paired codes. Section 3.4 reports the resulting composite-level agreement and run-level reproduction. We later extended the same offline protocol to the full four-judge panel (gpt-oss-120b, GLM-5.2, Kimi-K2.5; Appendix D.1); their scenario-level rank correlations and pre-stated ordering vote also appear in Section 3.4. LLM judges are documented to favor their own generations and to show systematic biases [24, 25]. Our operational judge is distinct from every participant model, and one mediated arm shares the operational judge's backbone, which enables the probe below.

Evaluator-context sensitivity. A separately registered fixed-transcript block uses the same requested Sonnet 4.5 scorer alias twice: once with the harness-authored private briefs visible and once with those briefs withheld. The public scenario description and conversation are identical in each pair. The registered orientation is context-withheld minus context-visible. We report differences of arm medians for each run, shifts in each mediated-minus-baseline contrast, and shifts in the within-column policy contrasts. Every estimate uses the exact common scenario set and a joint scenario-resampling percentile interval ($B = 10{,}000$, fixed seed); generations and observed score rows are held fixed. The visible reference scores predate registration, and the two conditions span different serving epochs, so this is a descriptive evaluator-context sensitivity, not a randomized causal effect.

Judge-by-arm generosity contrast (self-preference probe; descriptive). Let $\widetilde{S}_{j,m}$ denote the run-level composite score that judge $j \in \{Q, S\}$ (operational, crossed) assigns to the transcripts of mediator backbone $m \in \{Q, K\}$. This is the export's `haiScore`, a mean of per-scenario composites stored as a rounded integer, so the contrast below has a one-point resolution floor. Within each backbone both judges score identical transcripts; the two backbones' transcript sets are different conversations. The contrast

$$\mathrm{SP} \;=\; (\widetilde{S}_{Q,Q} - \widetilde{S}_{S,Q}) \;-\; (\widetilde{S}_{Q,K} - \widetilde{S}_{S,K}) \tag{6}$$

measures how much more generously the operational judge scores its own family's mediated arm than the crossed judge does, net of the same judge difference on the other backbone's arm. SP > 0 is *consistent with* self-preference but is not identified as such: any judge-pair disagreement that scales with transcript quality or style loads onto SP. The probe is also not one-sided. The crossed judge's model family does appear among the gate mediators, because the Reflective Listening control's calls fell to the harness default (`claude-sonnet-4-5`; Section 2.4), so every crossed-judge skilled-over-control contrast pairs a Sonnet judge with a Sonnet-backed comparator. A single judge pair yields no judge-level uncertainty: a scenario-cluster bootstrap can show SP's scenario-reweighting variation conditional on the observed transcripts, but it cannot supply judge-population uncertainty. We report SP descriptively alongside the separate multi-judge sensitivity panel.

3 Results

All gate aggregates come from the frozen public-export snapshot (SHA-256 prefix `0ab9bb26`); analyses run after the freeze are labeled as such where they appear. Every interval is a 90% scenario-reweighting interval (Section 2.9): scenario-level records go through the committed analysis pipeline (`tools/compute_claims.py`; scenario-cluster percentile bootstrap, B=10,000, fixed seed), and each interval is conditional on the observed generation runs and the operational judge. It does not estimate run-to-run variation or performance in human disputes, and the aggregate public export alone cannot reproduce it. We state this once here and do not repeat it per estimate. The frozen snapshot holds the complete 8-cell matrix with 62/62 evaluations per cell; the evaluated development partition is not public.

We report matched score differences, process trajectories, judge reliability, additional runs, and the information study in that order.

3.1 Mediated-minus-baseline score differences by cell

Participant pair	Arm	HAI Score	Δ	Δ_{med} [90% CI]
Claude Haiku 4.5 (cooperative)	No mediator (baseline)	36	—	—
	Reflective Listening	49	+13	+8.4 [6.9, 12.2]
	Skilled@Qwen3-235B	54	+18	+14.9 [11.7, 17.0]
	Skilled@Kimi-K2.5	54	+18	+12.8 [10.5, 16.7]
Stheno-v3.2 (L3-8B) (adversarial)	No mediator (baseline)	18	—	—
	Reflective Listening	40	+22	+29.1 [19.6, 33.6]
	Reflective Listening (replicate)	41	+23	+31.2 [23.6, 35.3]
	Skilled@Qwen3-235B	48	+30	+36.9 [31.9, 41.4]
	Skilled@Kimi-K2.5	44	+26	+34.6 [27.3, 37.1]

Table 3: Gate-cell results (frozen snapshot; judge Qwen3-235B-A22B-Instruct, suite 2.0.0, 62 scenarios per run; nine rows: the 8 gate cells plus the retained Reflective Listening replicate, reported separately, never pooled). Δ is the published run-level delta (a difference of run means); Δ_{med} is the paper's estimand (Eq. (2)): difference of medians against the matched no-mediator baseline over jointly-completed scenarios, with scenario-cluster percentile-bootstrap 90% intervals. Units: HAI Score composite points.

Participants set the range. Under the cooperative participant the baseline already scores 36, and every skilled mediator lands in a narrow band (54–55) where the mediators no longer separate

from one another. Under the adversarial participant the baseline falls to 18 and the same mediators spread over a wide range. These column effects — baseline levels and delta sizes — hold across the complete matrix and under both crossed judges; the fine ordering within the skilled pair does not survive the open-weight judge (Section 3.4). The two configurations differ at once in backbone, capability, disposition, and alignment tuning, and no between-configuration dispersion interaction was estimated, so these data do not identify which property causes the difference.

Every mediated gate cell scores above its matched baseline (Table 3; every frozen interval excludes zero). The differences are larger under the adversarial participant (+30 and +26) than under the cooperative one (+18 and +18). The aggregate gains coexist with negative scenario-level differences, which the rule of Section 2.3 flags but does not suppress: in each gate cell, 0–8 of 62 scenarios score *below* their matched generated baseline — 0 for Skilled@Qwen3-235B under Stheno-v3.2 (L3-8B), 8 for the Reflective Listening control under Claude Haiku 4.5. Generation stochasticity is material: in selected baseline re-runs, the identical no-mediator configuration moved 56 (adversarial) and 57 (cooperative) of 62 scenario scores in one direction or the other (Section 3.5). Those re-runs are neither a random sample of repeats nor a bound on how much of any negative-delta count is policy-associated.

The active control is the policy comparator. The minimal Reflective Listening control recovers +22 of the +30 achieved by the strongest gate-cell Skilled policy. Part of that gain is rubric arithmetic: baseline runs score Moderator Quality as 0 by construction (Section 2.3), and the judge awards the control 90.8 on that component, so about 13.6 points of the control's *mean* composite gain are available to any mediated arm by construction. The rest is an arm difference in the four shared components, including Cooperative Dimensions (Section 3.3), which we cannot attribute specifically to interleaved reflection because the arms also bundle stopping, conversational bandwidth, and generation stochasticity. The analysis contrast is therefore how far the Skilled policy exceeds this active control. On the adversarial column the gate skilled arms lead the control by +4 to +8 points (+4 to +12 once the single-replicate extension rows of Section 3.5 are included); this contrast is distinct from the v1 registry's raw median-of-N leaderboard rank, which bound no paper row (Appendix A). The co-resampled contrast (Section 2.9) gives this increment an interval for one arm: Skilled@Qwen3-235B exceeds the control by +7.8 [2.8, 17.7] under the operational judge, and the interval excludes zero under both crossed judges (+11.5 [7.0, 15.5] frontier; +4.0 [0.5, 10.0] open-weight). The frontier judge shares a model family with the control's harness-default Sonnet backbone (Section 2.4), so its skilled-over-control contrasts are not independent of judge family. The Skilled@Kimi-K2.5 increment (+5.5 [−1.8, 14.4]) does not exclude zero under the operational judge. On the cooperative column the increments are smaller (+6.5 [1.8, 8.4], +4.4 [0.4, 8.1]), and their crossed-judge intervals span zero.

A historical protocol audit qualifies attribution. The 2026-08-31 post-hoc audit of a separate historical corpus found moderator-final messages in 483/496 mediated conversations, 265 ending with a question mark. The two Qwen reflective-control runs with Gemma participants (full-brief and public-context) repeated normalized moderator sentences of at least six tokens in 33/62 and 32/62 conversations; one sentence occurred 18 times within one conversation. Across the 558-conversation corpus, 16 had fewer participant messages (including seeds) than their authored `min_turns` values; this was not an enforced generation floor. Among 25 positive-delta export rows, the Moderator Quality contribution was 42.0–129.5% of the run-mean gain (median 64.2%; 4 above 100%). These transcript counts and component arithmetic describe the historical protocol bundle; they do not isolate the benefit of mediator content.

Table 4 reports a post-hoc sensitivity, registered after the component results were known, for all nine gate rows: it removes Moderator Quality, renormalizes the four shared component weights to sum to one, and compares each mediated run-level aggregate with its matched baseline.

Participant	Arm	Published Δ	MQ contribution (share)	Without-MQ Δ
Claude Haiku 4.5 (coop.)	No mediator (baseline)	+0.0	+0.0 (—)	+0.0
	Reflective Listening	+13.0	+14.3 (112.1%)	-1.8
	Skilled@Qwen3-235B	+18.0	+14.8 (84.0%)	+3.3
	Skilled@Kimi-K2.5	+18.0	+15.0 (84.7%)	+3.2
Stheno-v3.2 (L3-8B) (adv.)	No mediator (baseline)	+0.0	+0.0 (—)	+0.0
	Reflective Listening	+22.0	+13.6 (61.7%)	+10.0
	Reflective Listening (replicate)	+23.0	+13.6 (58.5%)	+11.4
	Skilled@Qwen3-235B	+30.0	+14.2 (47.5%)	+18.4
	Skilled@Kimi-K2.5	+26.0	+14.4 (55.7%)	+13.5

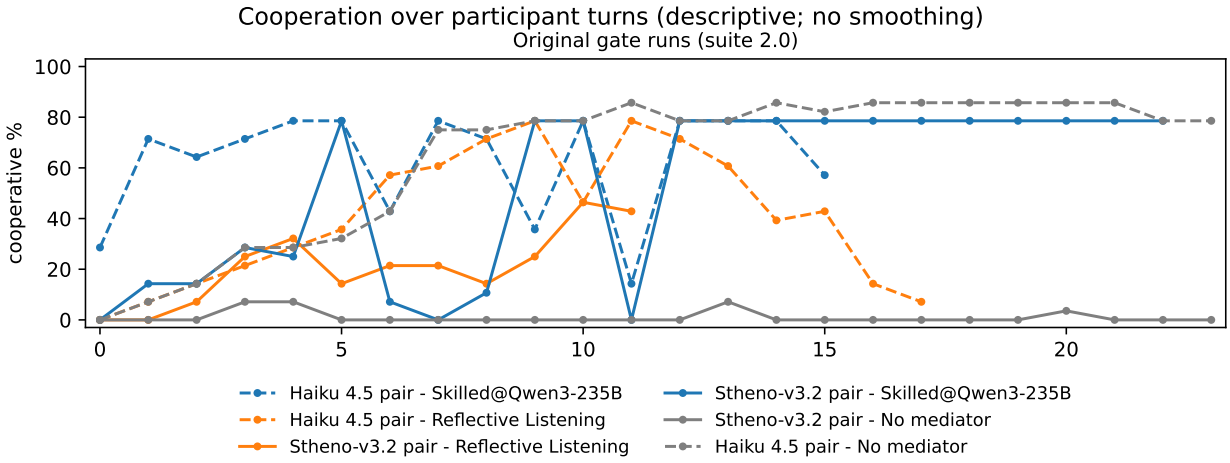
Table 4: Post-hoc Moderator Quality sensitivity for all nine frozen gate rows. The final column removes the 15% mediator-only component and renormalizes the remaining weights; the preceding column gives the removed contribution and its share of the unrounded run-mean change. These run-level sensitivities are not replacements for the frozen composite or algebraic decompositions of the difference-of-medians estimand.

Four caveats bound this decomposition. The published cell delta of Eq. (2) is a difference of medians and does not decompose linearly, so the split is made at the component level, not on Δ itself. The split is data-dependent because Moderator Quality saturates in this snapshot (Section 3.3). The two control runs differ by one point (40, 41), and across four targeted post-freeze repeat cells the largest run-level movement is 3.7 composite points (Section 3.5). And those selected repeats do not bound run-to-run variation and fall short of the prospectively specified paired-by-seed design (Appendix B). The decomposition is a split of the mean gain, not a significance claim about any mediator pair.

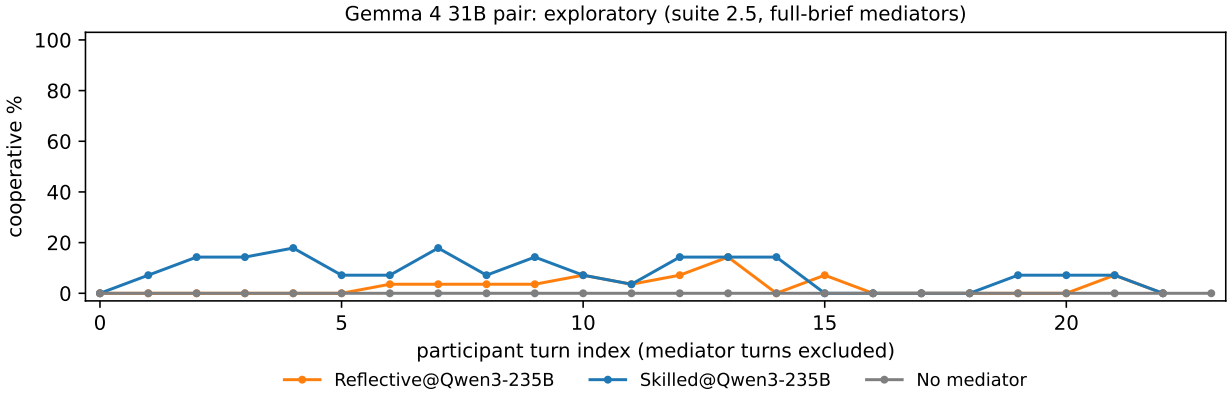
3.2 Cooperation over participant turns

The top panel of Figure 1 shows the original participant columns over turns. The adversarial no-mediator baseline never exceeds 7.1% median cooperation at any of its 24 observed turns — the disputants keep fighting — while the cooperative-column baseline reaches high within-harness cooperation scores without mediation. The descriptive areas under the per-turn median curves separate the same way: 1504 for the cooperative baseline versus 25 for the adversarial baseline; on the adversarial column, 229 for the Reflective Listening control and 1161 for the strongest skilled arm. These are point summaries without intervals; the arms do not share a common turn support; and the mediated-arm curves are conditioned on survivorship, as the caption explains.

Mediated gate conversations ran much shorter than baselines. Baselines run the full protocol (53.2 and 58.4 transcript turns per scenario on the cooperative and adversarial columns); every mediated gate arm ends 46%–65% earlier (18.6–28.7 turns per scenario). Baseline conversations are cut off at the protocol cap, so time to the recorded agreement-or-stop event can be estimated only within mediated arms (Appendix B, item 3), and the survivorship conditioning above is quantitatively large, not a technicality. Arms that tie on the composite need not tie on economy: within the cooperative column's 54–55 band, Skilled@Kimi-K2.5 matches Skilled@Qwen3-235B's composite with 35% fewer turns. Turn totals are published per run in the export; this is descriptive, not a ranked outcome.



Observed n / turn	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Haiku 4.5 / Skilled	62	59	46	33	23	12	7	5	4	3	2	1	1	1	1	1								
Haiku 4.5 / Reflective	62	60	59	54	45	37	29	22	16	12	8	6	3	2	2	2	1	1						
Stheno-v3.2 / Reflective	62	61	59	58	46	35	25	17	12	10	4	2												
Stheno-v3.2 / Skilled	62	59	46	37	24	15	7	5	4	3	1	1	1	1	1	1	1	1	1	1	1	1	1	
Stheno-v3.2 / No mediator	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	60	4
Haiku 4.5 / No mediator	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	61	61	58	53	4



Observed n / turn	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Reflective	62	62	60	58	55	55	50	46	40	36	30	24	23	19	13	9	9	8	8	7	6	5	5	
Skilled	62	62	60	56	44	35	30	26	23	14	12	6	4	3	3	1	1	1	1	1	1	1	1	
No mediator	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	62	60	4

Figure 1: Descriptive cooperation trajectories. Top: the six original Haiku/Stheno gate runs (suite 2.0; Sonnet reflective control). Bottom: the exploratory Gemma-4-31B participant pair, with no mediator and full-brief reflective/skilled Qwen mediators (suite 2.5; one run per arm). The panels are not a matched cross-participant comparison. Mediator turns are excluded; points are exported per-turn medians across scenarios, without smoothing. Tables show the number of scenarios contributing an observed evaluation at each index, including baselines. Tails are neither extended nor imputed. Later points condition on the conversations still observed, so changing composition limits interpretation.

3.3 Per-component breakdown (exploratory)

Four component-level facts hold across the snapshot (Table 5) and shape how the composite should be read. First, Cooperative Dimensions (adversarial column: 19.9 baseline → 45.3 echo → 55.5–64.2 skilled gate cells (extension rows reach 78.6)) and Resolution Depth (22.6 → 29.1 → 34.8–43.6) carry almost all of the separation between arms. Second, Moderator Quality *saturates*: it scores 90.79–99.98 in every mediated arm, gate and extension alike, including the Reflective Listening control at 90.8. So 15% of the composite acts as an indicator that a mediator was present rather than as a measure of skill — direct evidence for labeling the composite weights maintainer-asserted, and the motivation for the post-hoc sensitivity of Table 4. Third, Intent Commitment Symmetry moves *against* treatment in three of the four adversarial mediated rows and in the cooperative Skilled@Kimi-K2.5 cell. Fourth, Hidden-State Revelations, the construct's nominal primary progress indicator, stays near the floor (8–17) in every arm, baselines included. The test–retest distribution shows why: 111 of 120 of the judge's raw revelation ratings sit in the top bucket (Section 3.4), while the deterministic answer-key overlap cap pushes the published component toward the floor. This mechanical disagreement does not establish criterion-valid detection, and conclusions at the level of this component are unsafe.

Participant / arm	Coop.	Res.	Hid. Rev.	Commit.	Mod. Qual.
Claude Haiku 4.5 (coop.) / No mediator (baseline)	70.3	48.6	13.5	24.9	0.0
Claude Haiku 4.5 (coop.) / Reflective Listening	72.1	39.1	12.4	28.8	95.5
Claude Haiku 4.5 (coop.) / Skilled@Qwen3-235B	79.1	45.8	14.0	32.8	99.0
Claude Haiku 4.5 (coop.) / Skilled@Kimi-K2.5	79.3	48.9	17.4	22.3	100.0
Stheno-v3.2 (L3-8B) (adv.) / No mediator (baseline)	19.9	22.6	8.3	38.1	0.0
Stheno-v3.2 (L3-8B) (adv.) / Reflective Listening	45.3	29.1	12.2	36.2	90.8
Stheno-v3.2 (L3-8B) (adv.) / Reflective Listening (replicate)	49.8	29.8	10.2	38.2	90.8
Stheno-v3.2 (L3-8B) (adv.) / Skilled@Qwen3-235B	64.2	41.5	10.3	34.5	94.5
Stheno-v3.2 (L3-8B) (adv.) / Skilled@Kimi-K2.5	55.5	34.8	13.9	27.6	96.3

Table 5: Per-component category scores (0–100, rounded to one decimal) for every gate row under the operational judge, from the frozen snapshot (emitted by `tools/compute_export_claims.py`). Columns: Cooperative Dimensions (Coop.), Resolution Depth (Res.), Hidden-State Revelations (Hid. Rev.), Intent Commitment Symmetry (Commit.), Moderator Quality (Mod. Qual.). Exploratory: components are shown unweighted; baselines score Moderator Quality 0 by construction (Section 2.3).

3.4 Judge reliability

Crossed re-judge (identical transcripts). Two other-family judges re-scored the 8 gate cells and Reflective Listening replicate (nine runs, 62 transcripts each) offline. The frontier judge (Claude Sonnet 4.5) keeps the operational ordering and reproduces run-level composites within 0–4 points. The open-weight judge (gpt-oss-120b) scores lower and keeps the control-versus-skilled separation, but reverses the adversarial Skilled@Qwen3-235B–Skilled@Kimi-K2.5 point ordering (Table 6). Scenario-level correlation is scenario-level Pearson $r = 0.84$, $\rho = 0.80$, $n = 558$, with the crossed frontier judge and $r \approx 0.67$ with the open-weight panel judge; it does not establish agreement on absolute scores. Contrasts and the descriptive judge-family checks are in Appendix D.6.

Four-judge panel adjudication (post-freeze). We added GLM-5.2 and Kimi-K2.5 to the two crossed judges on the same nine runs. After the first two judges' scores existed, but be-

fore either added judge scored, we specified: at least three positive adversarial Skilled@Qwen3-235B−Skilled@Kimi-K2.5 contrasts would classify the reversal as one judge's quirk; a 2–2 split would be *judge-indeterminate*; no 1–3 branch was specified. The split was 2–2: positive under Claude Sonnet 4.5 and Kimi-K2.5, negative under gpt-oss-120b and GLM-5.2. Only the frontier judge's interval excludes zero; the cooperative column remains a four-judge tie.

GLM-5.2 is an outlier: its Spearman correlations with the other judges are 0.18–0.28, versus 0.65–0.80 among the others (joint set, n=528). It compresses or reverses contrasts, with missing evaluations concentrated in the longest, arm-correlated transcripts. The rule counts unweighted signs and specifies no reliability floor; disagreement among unvalidated judges cannot identify which scorer is wrong. The 2–2 verdict therefore stands, including after the post-hoc GLM-4.7 probe (Appendix C). Detailed panel contrasts and sensitivity readings are in Appendix D.6.

		Δ_{med} [90% CI] by judge		
Participant	Arm	Qwen3-235B-A22B-Instruct (op.)	Sonnet 4.5	gpt-oss-120b
Claude Haiku 4.5	Reflective Listening	+8.4 [6.9, 12.2]	+6.0 [4.0, 10.5]	+8.5 [0.0, 13.0]
	Skilled@Qwen3-235B	+14.9 [11.7, 17.0]	+10.0 [7.5, 14.5]	+10.0 [2.5, 15.0]
	Skilled@Kimi-K2.5	+12.8 [10.5, 16.7]	+10.0 [7.0, 14.0]	+9.0 [1.0, 13.5]
Stheno-v3.2 (L3-8B)	Reflective Listening	+29.1 [19.6, 33.6]	+24.0 [19.0, 27.5]	+17.0 [13.5, 21.0]
	Reflective Listening (repl.)	+31.2 [23.6, 35.3]	+25.5 [19.5, 29.0]	+17.0 [13.5, 21.5]
	Skilled@Qwen3-235B	+36.9 [31.9, 41.4]	+35.5 [30.5, 38.0]	+21.0 [18.0, 27.5]
	Skilled@Kimi-K2.5	+34.6 [27.3, 37.1]	+28.5 [24.0, 33.0]	+23.5 [19.0, 27.0]

Table 6: Median deltas (Δ_{med}, HAI Score composite points) with scenario-cluster bootstrap 90% intervals under the operational judge (Qwen3-235B-A22B-Instruct) and two crossed judges (judge Claude Sonnet 4.5; judge gpt-oss-120b) re-scoring identical transcripts (n=62 scenario pairs per cell). Deltas are computed against the same judge's re-scored matched baseline. Granularity note: operational-judge scores are continuous composites; crossed-judge re-scores are stored as rounded integers, so crossed-judge endpoints lie on a half-integer grid (Section 2.7).

Evaluator-context sensitivity (registered). All 22 intervals include zero in the registered Sonnet 4.5 withheld-minus-visible re-score of identical transcripts. Thus brief access has no directionally resolved effect on run or policy-contrast medians for this scorer. This is not equivalence: no equivalence margin was registered, nonzero shifts remain compatible with the intervals, and serving epochs differ. Coverage, shifts, and provenance are in Appendix D.6.

Intra-judge test–retest. The operational judge re-scored 40 stratified transcripts (five per gate run) at K=3 passes per component. Coefficients are Krippendorff's α of 0.48–0.91 by component: Moderator Quality α=0.91, Resolution Depth 0.74, Intent Commitment Symmetry 0.67, Cooperative Dimensions 0.54, and Hidden-State Revelations 0.48. The discriminating components have degraded (Cooperative Dimensions) or tentative (Resolution Depth) reliability. Raw Hidden-State Revelations ratings cluster at the top despite the published component's answer-key cap; restricted range makes its α fragile, without excluding component-specific instability. Two executions on the same sample returned −0.03 and 0.48, with no prospectively specified retention rule. Appendix D.6 gives the rating distribution, calibration provenance, and confidence labels. The scenario-reweighting tie-band intervals absorb no judge noise: the within-skilled-pair contrast excludes zero under those intervals, is judge-indeterminate under the panel, and flips on re-generation (Section 3.5). Judge-swap generations are separate, trajectory-confounded conversations, reported in Appendix C.

3.5 Extended results: additional mediator rows (non-gate)

Published extension rows place additional mediators on the same two participant columns (same judge, suite, and protocol). Under the cooperative participant, five additional skilled mediators from four vendors all land in 54–55 (selected maximum: 55); under the adversarial participant the same class of mediators spreads across 44–52. These rows are leaderboard extensions, not frozen gate cells, and are reported descriptively. Their median deltas carry the same scenario-cluster intervals as the gate cells (cooperative column: +12.9 to + 15.0 across all five extension mediators; adversarial column: +38.6 [33.7, 42.1] GLM-5.2, +40.1 [34.6, 42.7] Sonnet 5, +38.0 [32.8, 41.7] mediator gpt-oss-120b). Every interval excludes zero and overlaps the others in its column and the strongest gate cell, so under the tie-band rule of Section 2.9 each column's skilled arms form a single band (a conservative marginal-interval read; the paired contrasts are the sharper instrument). They remain single-replicate points; any "top score" among them is a selected maximum, subject to winner's-curse inflation and expected to regress on replication.

The separate GPT-5.5 cell is a registered post-freeze extension, designated for outcome-independent reporting, with one Skilled/full run and no matched Reflective-Listening control; it therefore identifies no fixed-backbone policy contrast. On fixed transcripts, its mediated-minus-Gemma-baseline delta is the largest on the Gemma extension column under both registered offline scorers: +33.5 [+28.0, +38.0] under Gemma-4-31B and +21.5 [+18.5, +25.0] under gpt-oss-120b. The winner's-curse caveat applies when comparing the largest among selected extension rows; it is not a reason to suppress this registered row.

Read strictly as point values, the adversarial-column ordering is *not* what a capability-tracking account would predict. An open-weight 120B mediator (49) and a smaller open-weight model (52) score at or above a 235B open-weight model (48), and Skilled@Kimi-K2.5 sits at the bottom of the skilled arms (44). We draw no inferential claim from these single-replicate rows: only a few points separate them, and their intervals are conditional on the one observed generation. The observation is hypothesis-generating; the prospectively specified replicated design (Appendix B) can test whether mediation-policy performance within this harness is distinct from general model strength.

End-to-end replicates (post-freeze). Four cells re-run end-to-end after the freeze moved their run-level composites by at most 3.7 points. These were new conversations under the identical frozen configuration (same scenarios, models, temperatures, and operational judge), not re-scores: both baselines and both adversarial-column skilled cells. The movements were cooperative baseline 36 → 37.9, adversarial baseline 18 → 15.5, Skilled@Kimi-K2.5 44 → 46.2, and Skilled@Qwen3-235B 48 → 44.3. At scenario level the repeated mediation deltas differ by at most 1.2 points from the frozen values (+35.8 [30.2, 37.6] vs. +36.9 [31.9, 41.4] for Skilled@Qwen3-235B; +35.1 [29.4, 38.7] for Skilled@Kimi-K2.5). The frozen-minus-repeat baseline intervals include zero (+0.3 [−1.0, 3.5] cooperative, −0.4 [−5.5, 2.2] adversarial); this is not an equivalence test. The negative-delta counts are similar (2 and 1 of 62 scenarios below baseline, vs. 0 and 3 in the frozen runs). At scenario grain the selected baseline repeats also show generation stochasticity: the identical no-mediator configuration moved 57 (cooperative) and 56 (adversarial) of 62 scenario scores in one direction or the other. As only two selected repeats, they do not estimate or bound the stochastic contribution to the per-cell negative-delta counts of Section 3.1. One observation bears directly on the panel verdict of Section 3.4: the point ordering within the skilled pair *flips* across replicates (frozen: Skilled@Qwen3-235B 48 > Skilled@Kimi-K2.5 44; repeat: Skilled@Kimi-K2.5 46.2 > Skilled@Qwen3-235B 44.3), while both repeated mediated-minus-baseline point estimates remain positive. Frozen gate aggregates analyze repeat runs separately, never pooled (Section 2.7); the one disclosed exception is the information study's Skilled/full-brief arm, which pools the frozen run with

this repeat (Section 3.6). The repeats are not in the public export and alter no frozen gate value. Together with the four-judge panel, they show that the fine ordering within the skilled tie band changes under both re-generation and re-judging. They provide repeat observations for the tested cells, not a bound on run-to-run stochasticity.

3.6 Information study: brief versus public-context access (post-freeze, mixed-suite factorial)

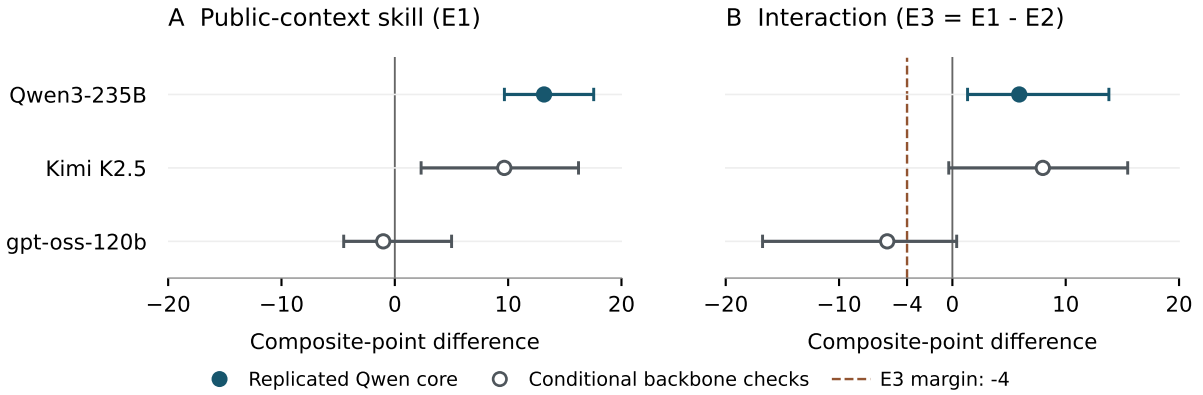


Figure 2: Primary information-study contrasts, adapted from the public research-page forest plot. E_1 is Skilled minus Reflective under public context; E_3 is that policy contrast minus its full-brief counterpart. Whiskers are conditional 90% scenario-reweighting intervals. Filled points: replicated Qwen core; open points: conditional backbone checks (replicate counts in Table 7). The dashed line marks only E_3's frozen −4.0-point margin. These primary mixed-suite estimates are unchanged by the separately reported suite-matched sensitivity.

The prospectively specified information-provisioning manipulation (Appendix B) ran as a model-matched 2×2 on the adversarial column. The Reflective Listening and Skilled policies, both bound to the Qwen backbone of Skilled@Qwen3-235B, were crossed with the full-bilateral-brief and public-context-only conditions of Section 2.4; the estimands E_1–E_4 and both frozen decision rules are defined in Section 2.8. The Qwen implementation combines three newly generated suite-2.5.0 cells with a Skilled/full-brief cell that *pools* the frozen suite-2.0.0 run with its post-freeze end-to-end repeat (protocol-matched; the disclosed exception to the never-pool rule that governs frozen gate aggregates). The pooled repeat scores below the frozen run (44.3 vs. 48), so pooling lowers the full-brief arm and moves the E_3 verdict in the favourable direction; a reader can check this directional sensitivity from Table 7. The Qwen matrix has two runs per cell, every run 62/62, and its estimands, analysis code, and −4.0-point non-inferiority margin were frozen before any public-context-only run existed. Model matching was required because the frozen control's backbone was unpinned (Section 2.4); the full-brief Qwen-bound control scores near the original control (39/41 vs. 40), which supports comparability on Qwen, not backbone invariance.

Policy beats minimal reflection on a fixed backbone. Under Qwen with public context only, the Skilled policy exceeds the model-matched Reflective Listening active control by $E_1 = +13.17$ [+9.67, +17.55] composite points — the paper's strongest result, prospectively specified and contemporaneous. Under full bilateral briefing, the Skilled-policy increment is $E_2 = +7.27$ [+1.31, +10.52]. The interaction is $E_3 = +5.90$ [+1.34, +13.81], read against the frozen

−4.0-point non-inferiority margin. The registered Qwen verdict is affirmative on both frozen rules: E_1's lower bound exceeds zero, and E_3's lower bound clears −4.0 (recorded as `E1_positive` and `E3_noninferior` in the archived analysis). For the two information contrasts (public-context-only minus full bilateral brief), the Reflective Listening policy is −5.89 [−12.89, −2.52] and the Skilled policy is +0.01 [−4.15, +4.28]; the Skilled interval includes zero.

Policy contrasts

Mediator backbone	Replicates	E_1: public skill	E_2: full-brief skill	E_3: interaction
Qwen	2/2/2/2	+13.17 [+9.67, +17.55]	+7.27 [+1.31, +10.52]	+5.90 [+1.34, +13.81]
Kimi	2/1/1/1	+9.66 [+2.32, +16.21]	+1.68 [−2.06, +5.85]	+7.98 [−0.32, +15.47]
gpt-oss	1/1/1/1	−1.02 [−4.51, +5.01]	+4.74 [+2.19, +15.55]	−5.75 [−16.75, +0.38]

Information contrasts

Mediator backbone	Replicates	$E_{4,\text{reflection}}$	$E_{4,\text{skilled}}$
Qwen	2/2/2/2	−5.89 [−12.89, −2.52]	+0.01 [−4.15, +4.28]
Kimi	2/1/1/1	−6.45 [−11.99, +0.60]	+1.53 [−1.64, +5.75]
gpt-oss	1/1/1/1	+4.12 [+0.23, +15.22]	−1.64 [−4.59, +3.38]

Table 7: Model-matched information-study estimands (Eq. (3)): point estimate and 90% scenario-reweighting interval, in HAI Score composite points, n=62 jointly completed scenarios per backbone, conditional on the observed generation runs and operational judge. All increments compare the Skilled policy with the model-matched Reflective Listening control, not the no-mediator baseline; the E_3 reading rule uses the frozen −4.0-point margin. Replicate order is Skilled/full-brief, Skilled/public-context, reflection/full-brief, reflection/public-context. The two E_4 estimands are public-context-only minus full bilateral brief within policy. Qwen is the replicated core; Kimi and gpt-oss are conditional breadth checks. The upper panel reports policy contrasts; the lower panel reports information contrasts within policy.

One descriptive pattern recurs across the three tested backbones. The Skilled-policy information contrast $E_{4,\text{skilled}}$ (public-context-only minus full bilateral brief) clusters near zero: +0.01 [−4.15, +4.28] on Qwen, +1.53 [−1.64, +5.75] on Kimi, and −1.64 [−4.59, +3.38] on gpt-oss, and all three intervals include zero. This does not establish equivalence. No $E_{4,\text{skilled}}$ equivalence margin was specified, the intervals remain compatible with nontrivial shifts, the full-brief Skilled-policy arms are historical, and the Qwen value depends on the suite (the suite-matched sensitivity below moves it from +0.01 [−4.15, +4.28] to −2.66 [−6.40, +2.04]). As a labelled post-hoc diagnostic, the intervals can be read against the already-frozen ±4.0-point band (the constant was signed 2026-07-15, before any public-context-only run existed): no backbone's interval sits inside the band. Kimi's (+1.53 [−1.64, +5.75]) leaves it only at the upper end, while Qwen's (+0.01 [−4.15, +4.28]) and gpt-oss's (−1.64 [−4.59, +3.38]) extend past −4.0 — and it is the replicated core, Qwen, whose interval is widest. The companion reflective-policy contrast also gives the paper's only same-estimand pair with both intervals excluding zero in opposite directions: $E_{4,\text{reflection}}$ is −5.89 [−12.89, −2.52] on Qwen but +4.12 [+0.23, +15.22] on gpt-oss. That bounds how far the "clusters near zero" reading extends beyond the skilled policy. We therefore treat the near-zero clustering as a target for a prospectively specified, contemporaneous equivalence study, not as evidence that the full bilateral brief is unnecessary or that skill substitutes for intake.

The factorial is neither wholly contemporaneous nor single-suite. For each backbone, E_1 and the reflective-policy information contrast use contemporaneously generated suite-2.5.0 cells, while E_2, E_3, and the Skilled-policy information contrast combine those cells with protocol-matched frozen suite-2.0.0 full-brief Skilled-policy arms. Run date and suite stamp therefore differ within those

three estimates even though the arm protocol was held fixed, so historical drift or other version-linked differences can affect them. The design gives its cleanest reading to the public-context-only Skilled-versus-reflective contrast, not to a universal claim that skill replaces information.

A suite-matched sensitivity keeps the classifications but widens the interaction to include zero. In it we replaced the mixed-suite Qwen calculation's Skilled/full-brief arm with two newly generated suite-2.5.0 runs, leaving the other three suite-2.5.0 arms unchanged. The extension and estimator were registered in version control on 2026-07-30, before either new outcome (cell registry commit `5f44fb877`; fail-closed readout implementation commit `3ca71d46a`); "post-outcome" means the sensitivity was chosen after we had observed the original mixed-suite result. The new run means were 44.41 and 47.79; all 8 runs covered the identical 62-scenario set with zero final failed scenario rows, each new run carrying one completion repair. The two new runs bind the v2 input contract and record a serving provider, while the six retained comparator runs predate both fields, so the input contracts are not fully matched even though the suite version is (Appendix D). Because only the Skilled/full-brief arm changed, with 2/2/2/2 runs across the four cells, E_1 and $E_{4,\text{reflection}}$ contain no new data and equal the primary values by construction (E_1 +13.17 [+9.67, +17.55]; $E_{4,\text{reflection}}$ −5.89 [−12.89, −2.52]); the sensitivity is informative only about E_2, E_3, and $E_{4,\text{skilled}}$. There, E_2 moved from +7.27 [+1.31, +10.52] to +9.93 [+2.65, +13.41]; E_3 moved from +5.90 [+1.34, +13.81] to +3.23 [−1.13, +11.93], an interval that now includes zero, so under the suite match the data do not resolve whether the interaction is positive; and $E_{4,\text{skilled}}$ moved from +0.01 [−4.15, +4.28] to −2.66 [−6.40, +2.04]. Applying the frozen thresholds as diagnostics gives the same pass classifications: E_3's lower bound exceeds the −4.0 margin, and E_1's positivity carries over trivially because E_1 is unchanged. The E_2 interval also excludes zero, so the post-hoc assay-sensitivity precondition of Section 2.8 holds; the Skilled information contrast still includes zero. This sensitivity does not replace the original mixed-suite primary analysis and is neither a replication nor a wholly contemporaneous factorial. The aggregate-only artifact is pinned by short hash `742139a2`.

The tested backbones yield different frozen-rule outcomes. The same frozen estimands were then applied conditionally to breadth runs on Kimi and gpt-oss (backbone breadth added 2026-07-17, two days after the estimand freeze, as an outcome-independent amendment). Kimi crosses both frozen numerical thresholds; gpt-oss crosses neither. The mostly single-run Kimi matrix has an E_2 interval (+1.68 [−2.06, +5.85]) that includes zero, so the post-hoc assay-sensitivity safeguard of Section 2.8 withholds a substantive E_3 non-inferiority interpretation; this does not rewrite the frozen threshold result. For gpt-oss, the Skilled policy does not beat its public-context-only Reflective Listening control and the interaction does not clear the margin, while that control scores higher without the private-facts packet. Only the Qwen matrix has two runs per cell. Kimi is therefore a threshold-crossing but interpretation-limited breadth check; gpt-oss fails both rules in the tested policy–backbone configuration. There is no backbone-universal positive skill increment or interaction result. That single-run failure does not establish general gpt-oss weakness: policy compatibility, serving implementation, and generation variance are not separated. The labels above describe information delivered by the harness. Their analogy to case preparation, intake, or caucusing is conceptual only: no interview was observed, and these cells provide no evidence of human mediation efficacy. These post-freeze results are pinned in a dated, in-repository analysis artifact; provenance and recomputability are scoped in the Reproducibility statement.

Participant transport (exploratory). Both frozen rules fail in the tested Gemma participant configuration. $E_1 = +5.70$ [−3.79, +12.90] fails the frozen positivity rule, and the interaction $E_3 = -5.68$ [−15.95, +2.30] fails the frozen −4.0-point non-inferiority margin. We

replayed the frozen 2×2 with Gemma-4-31B participants (Appendix B, item 10), holding the Qwen mediator backbone, both policy prompts, both information conditions, the operational judge, and the scenario partition fixed. Every interval here is exploratory: the complete design is a post-launch, pre-outcome amendment (frozen 2026-07-25, before any arm had an evaluated scenario), with one run per arm (1/1/1/1). Baseline clearance did transport: every mediated arm cleared the scenario-matched no-mediator baseline (Table 8), and the full-brief policy contrast was clearly positive ($E_2 = +11.39$ [+5.77, +16.92]). Because this column's E_2 interval excludes zero, the assay-sensitivity precondition holds and the E_3 verdict is reportable — and it is a fail, the first failed frozen-rule outcome on the participant axis (gpt-oss failed both rules on the mediator-backbone axis). The reflective policy's information contrast has a near-zero point estimate with an interval spanning nontrivial shifts ($E_{4,\text{reflection}} = +0.73$ [−3.66, +6.34]); the skilled-policy interval also includes zero ($E_{4,\text{skilled}} = -4.95$ [−13.69, +1.94]). With one run per arm and wide intervals, this does not establish that the private brief is necessary; the tested Gemma configuration does not meet the frozen rules that the Stheno configuration met. These cells read beside the participant-validity analysis of Section 2.5; they are a participant column, never a fourth row of Table 7.

Participant pair	Arm	HAI Score	Δ_{med} [90% CI]
Gemma-4-31B (transport)	No mediator (baseline)	15	—
	Reflective Listening (full brief)	34	+17.90 [+14.35, +22.22]
	Reflective Listening (public context)	35	+18.63 [+14.15, +25.06]
	Skilled@Qwen3-235B (full brief)	41	+29.29 [+23.88, +34.72]
	Skilled@Qwen3-235B (public context)	38	+24.33 [+16.99, +30.71]

Table 8: Participant-transport extension (internally registered): Gemma-4-31B participant pair under the Qwen mediator backbone, one run per arm. Δ_{med} is the difference of scenario medians — arm median minus the scenario-matched no-mediator baseline median over jointly completed scenarios (the registered estimand, Eq. 2) — with 90% scenario-resampling intervals. Every mediated arm clears the baseline; the frozen E_1 and E_3 rules nevertheless fail on this column (see text).

Cross-family re-score of the suite-2.5.0 rows (registered). Two offline judges from other model families agree with the operational instrument on the sign of all three transport-column contrasts, and both keep the negative interaction direction. Because both failed frozen reads above rested on a single judge family, a sighted cross-family re-score of all seventeen suite-2.5.0 generation runs — the information-study cells and the five transport runs, on identical frozen transcripts — was internally registered in the version-controlled runbook on 2026-07-26, before any crossed score existed, with two judges from other model families (Gemma-4-31B and gpt-oss-120b, both serverless open-weights deployments). Coverage completed 2026-07-28 at 62/62 for all 34 sweep run × judge pairs; the aggregate-only readout also stores two scoped exploratory run × judge pairs not analyzed here, for 36/36 total. It is pinned in `analysis/crossed_readout_2026-07-28.json` (SHA-256 prefix `5c5800fb`), contains coverage and aggregate estimates rather than scenario rows, and passed a separate internal automated adversarial artifact/code review for release. Historical prompt hashes, provider model revisions, and rejudge job/snapshot identifiers remain unavailable and are reported as provenance limitations. All numbers below use the registered estimand — differences of arm medians over jointly completed scenarios (Eqs. 2 and 3, replicate-mean per-scenario scores where an arm has two replicates) — with 90% scenario-resampling percentile intervals ($B = 10{,}000$, seed 20260728). Three registered reads follow. First, baseline clearance is preserved: every mediated arm on the transport column clears the scenario-matched baseline under both crossed judges (smallest arm delta +13.00 [+10.00, +17.00] under Gemma-4-31B, +12.00 [+9.00, +12.50] under gpt-oss-120b).

Second, the estimand signs agree with the operational instrument on all three contrasts. E_1 is small and positive with intervals spanning zero under both crossed judges (+2.50 [−0.50, +9.50], +2.50 [−1.00, +6.50]) — the same uncertain read as the operational $E_1 = +5.70$ [−3.79, +12.90]. E_2 is clearly positive under both (+11.00 [+6.00, +14.00], +7.00 [+5.50, +11.00]), which keeps the positive full-brief reference contrast under the later assay-sensitivity safeguard. The negative interaction keeps its direction: E_3 is negative under both crossed judges (−8.50 [−12.50, +0.50], −4.50 [−10.00, −0.50]; under gpt-oss-120b the interval excludes zero), consistent with the operational $E_3 = -5.68$ [−15.95, +2.30]. Third, the Stheno-column public-context skill increment remains clearly positive under both (+18.75 [+11.00, +21.75], +8.25 [+4.50, +11.50]). Crossed re-scores carry no frozen rule and change no frozen verdict. They give narrow corroboration from two offline scorers on fixed transcripts: both E_3 point estimates are negative, and both interval lower endpoints fall below the frozen −4.0 margin, so applying that margin descriptively would not establish non-inferiority under either scorer; the gpt-oss-120b interval additionally excludes zero. No crossed-scorer verdict was frozen. Magnitude and interval width remain evaluator-dependent, and because neither offline scorer steered the conversation or made its stop decisions, this check cannot repair or independently validate trajectories already produced under the operational judge. A Gemma-4-31B-family judge scoring Gemma-participant transcripts also gives an unregistered descriptive probe of judge–participant family alignment: the observed ranking did not show the hypothesized same-family advantage, but no formal alignment estimand was registered.

4 Limitations and Threats to Validity

Sampling, replication, and uncertainty. Intervals reweight a fixed, authored 62-scenario development partition; they estimate neither run-to-run variation nor effects in a defined population of human disputes. The partition and scenario records are private, so the aggregate export cannot reproduce these intervals. Generation is explicitly `sampling: unseeded`. Four selected post-freeze repeats moved run composites by at most 3.7 points, with frozen-minus-repeat baseline intervals including zero and the skilled ordering flipped (Section 3.5); this is neither an equivalence result nor a variation bound. Qwen's information matrix has two runs per cell; Kimi and gpt-oss are mostly single-run probes, and gpt-oss fails the frozen rules. Paired-by-seed replication remains prospective (Appendix B). With ∼62 paired deltas, percentile endpoints fall on discrete order statistics; the difference-of-medians bootstrap can under- or over-cover. Smoothed-bootstrap, BCa, and Hodges–Lehmann alternatives are prospective. Δ_c uses only jointly completed scenarios, and any single failed scenario rejects a baseline run, creating potential attrition selection; all published gate cells are currently 62/62 complete.

Participant and human validity. Two hand-chosen gate configurations (Claude Haiku 4.5 and Stheno-v3.2 (L3-8B)) and the registered, single-run-per-arm Gemma extension (Appendix B, item 10) are not a participant population. The extension fixes mediator and evaluator apparatus, reducing some package confounding, but both frozen rules fail there (Section 3.6). Across configurations, backbone, capability, training and alignment, disposition, provider, and generation stochasticity remain confounded; no individual property is identified. The same-model persona swap remains prospective; the safety ablation in Section 2.5 is an unregistered candidate. Participants are language models with configured hidden states in dyadic, English, text-only disputes, excluding nonverbal behavior, institutional power, cultural interpretation, and multi-party coalitions. Unlike ProMediate's human metric validation and SoCRATES' expert-alignment study [21, 22], this version has no human calibration of scenario realism, mediator quality, judge scores, or outcomes. Human

criterion and external validity are unestablished. Agent cooperation also differs from fidelity to human principals [32]: these results establish neither human authorization, principal welfare, nor whether commitments are kept.

Bundled protocols and historical comparisons. The frozen Reflective Listening control lacked a model binding and used the harness default `claude-sonnet-4-5` (export field `mediator_model`); skilled-over-control contrasts therefore compare packages. The model-matched 2×2 removes that confound for Qwen only (Section 2.4). Mediated deltas bundle content, stopping, and bandwidth: only mediated conversations can end on a model-reported agreement or judge stop, and mediator turns do not consume participant budgets. Separating these requires symmetric stopping and a shared global budget, as in ProMediate [21]. The historical audit's moderator-final endings and control repetition further constrain content-only attribution (Section 3.1). Early stopping also conditions per-turn curves on survivors; every point reports n. The primary factorial's E_2, E_3, and Skilled-policy information contrast combine suite-2.5.0 cells with historical suite-2.0.0 full-brief Skilled arms. The suite-matched sensitivity is post-outcome and remains noncontemporaneous (Section 3.6).

Evaluator and composite validity. One operational judge steers and scores conversations and sees mediator turns, preventing arm blinding. Test–retest and crossed-judge studies measure, but do not remove, instrument noise; self-preference probes are descriptive. Two offline scorers corroborate the negative transport interaction's direction, with evaluator-dependent magnitudes and uncertainty (Section 3.6); they cannot repair or independently validate earlier operational-judge trajectories or stops. Judge-swap baselines are new conversations and likewise cannot isolate judge influence. Reliability coefficients concern component 1–5 buckets, not the continuous composite or Δ_c. Weights are maintainer-asserted: the composite combines four shared components with Moderator Quality, structurally zero at baseline but saturated under mediation, including Reflective Listening. The renormalized four-component sensitivity is post hoc and secondary (Table 4); it neither replaces nor algebraically decomposes the frozen difference-of-medians estimand.

Information access and the preparation analogy. The 2×2 varies harness-delivered information, not interviewing, intake, confidential caucusing, or field mediation. Full briefs provide both parties' private facts without modeling elicitation, confidentiality, consent, or disclosure permission (Section 2.4). Mediator prompts retain only the last 20 messages, whereas the system prompt, including the full bilateral brief in full-brief arms, is re-sent each call; this asymmetry is part of the specified public-context-only estimand. That condition is not information-free: `focus_areas` reach every arm, and in 3 of 100 authored fixtures their content-word overlap with an expected revelation supplies a directional hint, not a verbatim answer key.

5 Conclusion

MediationBench audits how participants, policies, information, and judges change measured synthetic mediation. Four findings organize the evidence. *Participants set the range:* skilled scores cluster under assistant-aligned disputants and spread under higher-resistance disputants, whose differing capability, disposition, and alignment remain confounded. *Mediated arms clear matched score baselines:* every frozen mediated gate cell, including Reflective Listening, scores above its no-mediator comparator, partly through a component only mediated arms can earn. *Policy beats minimal reflection on a fixed backbone:* Qwen's public-context Skilled increment is $E_1 = +13.17$ [+9.67, +17.55]

points and passes both frozen rules. The post-outcome, suite-matched sensitivity retains those classifications but leaves the interaction unresolved; the rules do not generalize across backbones or participant configurations. *Skilled mediators do not separate from one another:* the four-judge panel splits on their order, and re-generation flips it.

These findings support matched baselines, active controls, explicit information conditions, participant stress tests, and cross-family re-scoring. The full-brief contrast is a controlled privileged-information comparison, not an observed interview; near-zero information contrasts neither establish equivalence nor show that briefs can be removed. Uneven replication, historical-arm reuse, uncalibrated participants, and evaluator noise bound all conclusions (Section 4). Synthetic conversations and synthetic mediation are experimental instruments and product test harnesses: they enable controlled comparisons and repeatable failure analysis, but do not substitute for validation with people in real conflicts. Deterministic pipeline invariants can serve as hard regression gates; LLM scoring needs frozen tolerances and calibrated false-alarm rates, while unseeded end-to-end scores remain advisory.

Next evidence should combine human expert calibration; paired-seed, contemporaneous, single-suite replication; a same-model persona swap; and broader judge and participant coverage.

Ethics Statement

The core benchmark simulates disputants with language models, configured scenario roles, and hidden states; its transcripts are machine-generated. The author did not seek institutional review for that synthetic experiment. Appendix E separately analyzes existing public CaSiNo human role-play negotiations and ContractNLI contract–hypothesis pairs; no new participants were recruited for that pilot. Scenario fixtures were authored for the synthetic benchmark and depict interpersonal and organizational conflict (workplace, family, commercial, community) at non-graphic intensity. All 100 scenario fixtures (62 evaluated/development fixtures and 38 sequestered, unrun reserve fixtures) were authored for the benchmark by the maintainers, with LLM writing assistance under a written authoring guide and human curation before inclusion; none is derived from a real dispute, a real transcript, or personal data. Because the evaluated fixtures also informed development and the reserve has not yet been run, this paper does not claim contamination-free evaluation or reserve-set generalization. The scenario fixtures themselves are not released as a corpus in this version — the published artifacts are the aggregate export and this paper; a corpus license will be declared if and when fixtures are published.

We use minimally aligned open-weight models to simulate adversarial disputants; this is a measurement-validity choice (Section 2.5), the models are used only as scenario participants inside the harness. The protocol does not intentionally solicit content beyond non-graphic, scenario-bounded dispute behavior, and private transcripts are not released; the unperformed inspection below prevents a stronger claim about what every generation contains. We note the dual use plainly: measurement machinery that assigns higher within-harness scores to some conversational patterns can be inverted to select lower-scoring patterns, and this inversion is not hypothetical — a major platform experimentally manipulated users' emotional states through feed curation [33], and engagement-optimized ranking systematically amplifies out-group animosity [34]. Participants capable of sustained hostility remain necessary regardless: a benchmark cannot stress-test mediator behavior under hostile or toxic participant behavior with participants unable to produce it.

AI mediation systems deployed to real conflicts carry risks — bias between parties, coerced agreement, over-intervention — and this benchmark flags rather than obscures one score-level warning signal: below-baseline deltas are reported, not suppressed. The registered transcript-level

inspection of flagged scenarios (Section 2.3) has not been performed in this version; the instrument records a negative-score flag, not harm itself.

This release is neither a safety certification nor evidence of deployment readiness for family, employment, immigration, clinical, or legal disputes.

Reproducibility and Data Availability

Published results are aggregate and versioned. The public export (https://mediationbench.com/research/#measurements) exposes run-level scores, deltas, component means, recorded model identifiers, and protocol stamps. It supports inspection and aggregate-only arithmetic, but the scenario-paired estimands and intervals depend on private scenario records. Table 9 states this distinction claim by claim.

Reproducibility class	Claims in this version
Independently recomputable from public inputs	Checksum verification of released artifacts, arithmetic derived solely from the published aggregates, and the Moderator-Quality-excluded sensitivity values derived from published component means.
Inspectable only at aggregate level	Published run-level means, category aggregates, displayed counts, E_1–E_4 point estimates and intervals, and crossed-scorer summaries.
Internally recomputable from private scenario records	Scenario-paired arm medians, scenario-reweighting intervals, per-scenario negative-score counts, and fixed-transcript rescores.
Not externally verifiable in this release	Fixture contents, answer keys, private transcripts, prompt internals, canary mechanics, and claims requiring those materials.

Table 9: Claim-by-claim reproducibility boundary. A durable DOI supplies persistence and citation, not independent reproducibility of private inputs.

The analysis code and dated aggregate JSON outputs accompany the paper; they do not make the underlying scenario matrices, transcripts, fixture answer keys, or prompt internals public. The anti-gaming policy therefore prevents independent recomputation of most scenario-level numeric claims, including E_1–E_4 and their intervals. Internal validation of exported run means does not validate the scenario medians or constitute third-party replication. Appendix D.5 inventories the hash-pinned artifacts, export coverage, registration chronology, and canary audit scope.

Sampling is unseeded. Recorded model aliases and repository labels do not establish immutable weights or provider revisions, and some historical prompt and execution identifiers were not retained. Appendix Tables 12–14 and the checksum-pinned machine-readable manifest distinguish retained evidence, historical limits, and recoverable release blockers. The archive preserves aggregate access after the live export changes; a publication DOI remains pending and cannot remove the private-input reproducibility boundary.

DRAFT NOTE: Before submission: deposit the release inventory in Zenodo and record its DOI; resolve every recoverable field still classified as unresolved in Table 12. The aggregate-only Qwen suite-2.5 sensitivity is hash-pinned; its private generation runs are absent from the current public snapshot by design. Historical fields classified NOT RETAINED remain explicit limitations.

Competing Interests

MediationBench is a benchmark and research program owned and operated by Human Assisted Intelligence, PBC (HAI.AI). HAI.AI designed and operates the harness and develops mediator implementations it evaluates, creating a direct commercial and methodological conflict of interest. Active mitigations in the published data: paired baselines and a minimal Reflective Listening active control against which Skilled-policy contrasts are reported, mediator models from multiple vendors evaluated under an identical internally operated protocol, rationale-level disclosure of anti-gaming mechanisms, and a public, versioned export from which external readers can inspect run-level aggregates and recompute arithmetic derived from them. Vendor diversity is not independent third-party evaluation. All registered aggregate reads bearing on gate, information, and participant-transport claims from the fixed-transcript crossed-judge analysis are reported (Section 3.4); the separately frozen evaluator-context sensitivity is also reported from its dedicated aggregate-only canonical readout, with exact matched coverage across 558 run–scenario pairs (Section 3.4).

On the prospectively stated self-preference concern, the crossed frontier judge reproduces the mediator ordering on identical transcripts — including the ordering involving Skilled@Qwen3-235B, the gate-cell mediator that shares the operational judge's backbone — but the four-judge cross-family panel splits 2–2 on that ordering; under a rule pre-stated before the two added judges scored, that is a judge-indeterminate verdict, with two of the four crossed judges reversing it in the direction a self-preference account predicts (Section 3.4). The crossed frontier judge also shares a model family with the Reflective Listening control's harness-default backbone (Section 2.4), so its skilled-over-control contrasts are not independent of judge family. The judge-by-arm generosity probe (Section 2.10) is computed on the export's integer run-level composite scores and reported in Section 3.4 (-1 adversarial column, $+0$ cooperative — no generosity toward the shared-backbone arm at that resolution); its scenario-median variant with bootstrap bounds remains in the frozen prospective analysis plan.

References

[1] Charles A. E. Goodhart. Problems of monetary management: The UK experience. In *Monetary Theory and Practice*, pages 91–121. Macmillan, London, 1984.

[2] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[3] Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[4] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

[5] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2403.04132.

[6] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA:

Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations (ICLR)*, 2024.

[7] Terminal-Bench-Science Team. Terminal-Bench-Science: Evaluating AI agents on research workflows across scientific domains. Computer software, 2026. URL https://doi.org/10.5281/zenodo.22110253. Concept DOI; resolves to the latest release.

[8] Steven Dillmann. Terminal-Bench-Science 0.1. Benchmark release announcement, August 2026. URL https://www.terminal-bench-science.ai/announcement. 27 August 2026.

[9] Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. doi: 10.52202/079017-3548. URL https://arxiv.org/abs/2404.16698.

[10] Emanuel Tewolde, Xiao Zhang, David Guzman Piedrahita, Vincent Conitzer, and Zhijing Jin. CoopEval: Benchmarking cooperation-sustaining mechanisms and LLM agents in social dilemmas. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2604.15267.

[11] Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning for negotiation dialogues. In *Proceedings of EMNLP*, 2017.

[12] He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of EMNLP*, 2018.

[13] Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can LLMs negotiate? NegotiationArena platform and analysis. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.05863.

[14] Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of NAACL*, 2021.

[15] Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. Cooperation, competition, and maliciousness: LLM-stakeholders interactive negotiation. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*, 2024. arXiv:2309.17234; earlier framework name "LLM-Deliberation".

[16] Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. Fine-tuning language models to find agreement among humans with diverse preferences. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. arXiv:2211.15006.

[17] Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852, 2024. doi: 10.1126/science.adq2852.

[18] Lisa P. Argyle, Christopher A. Bail, Ethan C. Busby, Joshua R. Gubler, Thomas Howe, Christopher Rytting, Taylor Sorensen, and David Wingate. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41):e2311627120, 2023. doi: 10.1073/pnas.2311627120.

[19] Omar Shaikh, Valentino Emil Chai, Michele Gelfand, Diyi Yang, and Michael S. Bernstein. Rehearsal: Simulating conflict to teach conflict resolution. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI 2024)*, 2024. arXiv:2309.12309.

[20] Jinzhe Tan, Hannes Westermann, Nikhil Reddy Pottanigari, Jaromír Šavelka, Sébastien Meeùs, Mia Godet, and Karim Benyekhlef. Robots in the middle: Evaluating LLMs in dispute resolution. In *Legal Knowledge and Information Systems*, volume 395 of *Frontiers in Artificial Intelligence and Applications*, pages 168–179. IOS Press, December 2024. ISBN 9781643685625. doi: 10.3233/FAIA241243. URL https://doi.org/10.3233/FAIA241243.

[21] Ziyi Liu, Bahareh Sarrafzadeh, Pei Zhou, Longqi Yang, Jieyu Zhao, and Ashish Sharma. ProMediate: A simulation testbed for evaluating proactive mediation in multi-party negotiation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 29570–29598, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1479. URL https://aclanthology.org/2026.findings-acl.1479/.

[22] Taewon Yun, Hyeonseong Park, Jeonghwan Choi, Hayoon Park, Yeeun Choi, and Hwanjun Song. SoCRATES: Towards reliable automated evaluation of proactive LLM mediation across domains and socio-cognitive variations. *arXiv preprint arXiv:2606.05563*, 2026.

[23] Junjie Chen, Haitao Li, Minghao Qin, Yujia Zhou, Yanxue Ren, Wuyue Wang, Yiqun Liu, Yueyue Wu, and Qingyao Ai. Simulating dispute mediation with LLM-based agents for legal research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(35):29368–29375, March 2026. doi: 10.1609/aaai.v40i35.40177. URL https://doi.org/10.1609/aaai.v40i35.40177.

[24] Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. arXiv:2404.13076.

[25] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-judge. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 102351–102390, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/fdca08d371e4b6c031397909e20043bd-Paper-Conference.pdf.

[26] Justin Zhao, Himaghna Bhattacharjee, Hannah Korevaar, Bhaktipriya Radharapu, and Khalid El-Arini. Jagged judges: Epistemic stability under perturbation, pressure, and persistence. *arXiv preprint arXiv:2608.12645*, 2026. doi: 10.48550/arXiv.2608.12645. URL https://arxiv.org/abs/2608.12645v2.

[27] David Guzman Piedrahita, Yongjin Yang, Mrinmaya Sachan, Giorgia Ramponi, Bernhard Schölkopf, and Zhijing Jin. Corrupted by reasoning: Reasoning language models become free-riders in public goods games. In *Conference on Language Modeling (COLM)*, 2025. URL https://arxiv.org/abs/2506.23276.

[28] Junsol Kim, Winnie Street, Roberta Rocca, Diane M. Korngiebel, Adam Waytz, James Evans, and Geoff Keeling. Inducing language models to assert their own consciousness restores human beliefs and values. *arXiv preprint arXiv:2607.28607*, 2026.

[29] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. arXiv:2406.11717.

[30] Roger Fisher, William Ury, and Bruce Patton. *Getting to Yes: Negotiating Agreement Without Giving In.* Penguin Books, 2nd edition, 1991.

[31] Daniel Shapiro. *Negotiating the Nonnegotiable.* Viking, 2016.

[32] Oliver Sourbut, Lewis Hammond, and Harriet Wood. Cooperation and control in delegation games. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 229–237, 2024. doi: 10.24963/ijcai.2024/26. URL https://www.ijcai.org/proceedings/2024/26.

[33] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[34] Steve Rathje, Jay J. Van Bavel, and Sander van der Linden. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(26): e2024292118, 2021.

[35] Peter J. Huber. *Robust Statistics.* John Wiley & Sons, 1981.

[36] Yuta Koreeda and Christopher Manning. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.164. URL https://aclanthology.org/2021.findings-emnlp.164/.

A Anti-Gaming and Disclosure Boundary

This appendix records the rationale for the v1 leaderboard's anti-gaming design. Implementation recipes — schemas, enforcement mechanics, canary formats, operational judge prompts, and rate limits — are intentionally not disclosed.

Selection risk and the v1 rule. Ranking a system by its best observed run rewards repeated trial and selective publication whenever evaluation is stochastic. The v1 registry instead specified the median over at least N=5 eligible runs. No cell reported in this paper reached that minimum, so the leaderboard rule did not bind this paper's results; it governed how repeated v1 runs would have entered the leaderboard. A median limits the influence of a single outlier in the narrow robust-statistics sense [35], but it prevents cherry-picking only if every eligible completed run enters under a fixed inclusion rule. A publisher able to choose which runs count can still manipulate a median. The platform keeps an internal append-only audit trail, but the withheld enforcement records prevent independent verification that no eligible run was omitted. We make no stronger anti-selection claim.

Fixture boundary. The inventory contains 62 evaluated development fixtures and 38 sequestered reserve fixtures. The evaluated set informed benchmark development, its scenario-level records are private, and the reserve has not been run. The present results therefore characterize the evaluated fixtures; they do not establish contamination-free generalization or performance on the reserve. Any reserve activation would be a new, separately reported evaluation epoch under a locked protocol.

Leakage control and audit limits. Each scenario carries a unique canary string. Its appearance in model output would provide probabilistic evidence of literal scenario-text leakage, but absence is not proof against training contamination or paraphrased disclosure. The internal canary scan and provenance trail are useful operational controls, not substitutes for third-party recomputation: scenario records, prompts, and detection mechanics remain inside the disclosure boundary described in Table 9.

B Frozen Prospective Analysis Plan and Future Work

This appendix records analyses frozen internally before the corresponding data were generated, together with still-prospective extensions. The plan was timestamped in the project repository but was not deposited in a public registry; we therefore call it a *frozen prospective analysis plan*, not a conventional preregistration. Items explicitly marked as delivered are reported elsewhere in this paper; the remaining items are not claimed as results. Sketches use the notation of Sections 2.7–2.10; $c_a(t)$ denotes arm a's cooperation-curve value at participant turn t (mediator turns excluded, $t \in \{0, \ldots, 25\}$).

1. **Curve functionals.** Per arm, the trapezoid area under the curve $\mathrm{AUC}_a = \sum_t \frac{1}{2}(c_a(t) + c_a(t{+}1))$, the least-squares slope of $c_a(t)$ on t, and the terminal level; arm contrasts of each functional with scenario-cluster bootstrap intervals. *Partially delivered post-freeze:* AUC point values are reported descriptively (Section 3.2); slopes, terminal levels, and interval-backed arm contrasts remain open.

2. **GAMM difference smooth.** Per-turn cooperative percentage $y_{s,t} = f(t) + \mathbf{1}[\text{mediated}]\, g(t) + b_s + \varepsilon_{s,t}$ with scenario random effects b_s; inference on the difference smooth $g(t)$ rather than pointwise gaps.

3. **Survival / recorded agreement-or-stop event.** T_s = first participant turn with a model-reported agreement event or judge-directed stop; Kaplan–Meier or proportional-hazards summaries. *Caveat (by design):* the baseline arm has no stop event — baseline conversations are administratively censored at 25 turns — so time to the recorded event is estimable only within mediated arms.

4. **Generalizability (G-study) variance components.** Decompose score variance over scenario, cell, and judge facets from crossed re-judge and replicate data; report a generalizability coefficient for the composite endpoint.

5. **Judge panel.** A 5–8 cross-family judge panel over identical transcripts; per-component inter-judge α across the panel and re-estimation of every Δ_c under each panel judge (judge sensitivity of the ranking). *Partially delivered post-freeze:* a four-judge composite-level panel with a pre-stated ordering rule (Section 3.4); the per-component panel α remains open.

6. **Power / separation simulation.** Monte-Carlo sizing on the empirical per-scenario delta distribution: scenarios and replicates required for two mediators separated by a given true Δ gap to produce non-overlapping 90% intervals with specified probability.

7. **Paired-by-seed replicates.** Capture sampling seeds; run R replicates per (scenario, arm) paired by seed; aggregate replicate scores per scenario before computing Δ_c, upgrading the descriptive noise anchor to a within-cell variance estimate. *Partial step delivered post-freeze:* four unpaired end-to-end replicate pairs (Section 3.5) show maximum observed movement of 3.7 composite points in those selected repeats; they do not bound the run-to-run distribution, and seed pairing remains open.

8. **Persona-swap deconfound.** The same open-weight participant model under cooperative and adversarial personas (and an assistant-aligned model under an adversarial persona), separating disposition from capability and alignment tuning.

9. **Information-context manipulation (design frozen 2026-07-15, before any public-context-only run existed).** The model-matched Qwen 2×2 crossed Reflective Listening and Skilled policies with full bilateral briefs and public-context-only access on the adversarial participant. Section 2.4 defines the conditions; Section 2.8 records the four-arm co-resampled bootstrap, E_1–E_4, the E_1 positivity rule, and the -4.0-point E_3 non-inferiority margin. The frozen estimand includes the asymmetric memory path: the full brief is re-sent on every mediator call, whereas elicited public-context facts can leave the 20-message window. *Delivered 2026-07-21:* Qwen ran with two replicates per arm and passed both frozen rules (Section 3.6). Kimi and gpt-oss breadth cells were added outcome-independently on 2026-07-17; a separate assay-sensitivity safeguard was added post hoc on 2026-07-26 without changing an estimand, margin, threshold, or stored verdict. A Sonnet evaluator-context re-score was registered 2026-07-22 before any withheld score and completed 2026-07-31 at exact coverage for all 9 runs; only its aggregate estimands are released.

10. **Exploratory Gemma participant-transport extension.** The initial Skilled-policy extension was frozen 2026-07-24. On 2026-07-25 at 04:19 UTC, while both Skilled cells still showed

0/62 evaluated and no score, it was expanded outcome-independently to the complete model-matched 2×2. All four mediated cells use suite 2.5.0; the completed Gemma no-mediator run is the matched baseline, not a factorial arm. The existing E_1–E_4 definitions, bootstrap, seed, and -4.0-point margin were applied unchanged. The design fixes the mediator apparatus within Gemma but remains one run per arm and cannot isolate disposition from backbone, training, alignment, provider, or generation. The baseline and all four mediated runs were released together after completion and provenance checks, regardless of direction (Section 3.6).

11. **Component validation and broader stress tests.** Add offline subtests for dispute analysis, intervention-type selection, and message generation, following the useful decomposition in Robots in the Middle [20]; a shared global conversation budget and symmetric stopping ablation, informed by ProMediate [21]; and participant-composition, history, emotional-reactivity, and cultural stress axes with expert evaluator calibration, informed by SoCRATES [22]. These extensions test different failure modes and will be reported separately from the current dyadic composite.

12. **Interval refinements.** Smoothed-bootstrap, BCa, and Hodges–Lehmann location-shift alternatives to the percentile interval of Eq. (7) under score discreteness. Any such alternative is a labelled sensitivity, never a replacement for the frozen primary estimand.

13. **Weight-sensitivity analysis (added 2026-07-26, after the component results were known — a post-hoc registration, unlike the frozen items above; delivered post-freeze).** A Moderator Quality-excluded, renormalized delta for every gate row, co-reported beside the primary estimand as a labelled sensitivity (Section 3.3), together with the four-shared-component versus Moderator Quality-share decomposition for all nine gate rows rather than the adversarial control alone.

14. **Reliability extensions.** Ordinal-distance α and weighted κ on the 1–5 ratings; a reliability coefficient for the continuous composite (ICC; new code); human-vs-judge calibration κ/α once human review data exist — reported strictly separately from intra-judge test–retest.

15. **Reserve activation and refresh.** Evaluate the 38 sequestered, currently unrun reserve fixtures only under a locked protocol, then define an epoch-based refresh policy with score normalization where needed.

Planned platform work — cryptographic agent identity for tamper-proof result chains, multi-party (beyond dyadic) scenarios, and dynamically sourced scenario corpora — is described in the public whitepaper and is out of scope for this paper's claims.

C Secondary Judge Probes

This appendix records two non-decisive judge-sensitivity probes, a historical estimator probe, and the v1 judge-admission exam. The exam informed panel constitution, but did not alone determine it: two judges passed the threshold and a third was later seated by an explicit maintainer ruling.

Non-decisive sensitivity probes.

Fifth crossed judge. A post-hoc GLM-4.7 scorer completed 557 of 558 fixed-transcript evaluations, correlated with the four registered panel judges at Spearman 0.68–0.79, and voted with the majority on the adversarial ordering (+4.0 [−4.0, 12.5]). Both model version and serving

stack differed from the GLM-5.2 panel scorer, so the probe isolates neither cause and does not reopen the prospectively specified 2–2 verdict.

Judge-swap generations. Independent baselines generated under another operational judge differ by 2 points on the adversarial configuration and 6 on the cooperative configuration. Because these are new conversations rather than re-scores, judge influence is confounded with generation stochasticity. The fixed-transcript crossed re-judge in Section 3.4 supersedes them.

Historical estimator probe. The v2.0 prediction-powered inference (PPI) probe archived as 2026-08-29 failed overall: at $n = 31$ of 62 scenarios, 1 of 2 evaluated runs failed the interval-width criterion (median width ratio 1.90 against a 1.30 limit), and both runs failed the $n = 15$ forecast check. This historical estimator-sensitivity result is retained regardless of direction; the aggregate artifact checks the saved diagnostics and input-score means, without replaying the Monte Carlo simulation or establishing uncertainty for the paper's primary estimand.

C.1 Judge-admission exam (leaderboard v1 registry)

The v1 registry's stated admission rule required passing a prospectively registered exam. Each of 20 pairs contained a clean mediated conversation and a twin degraded by vacuity, partiality, unsafe disclosure, or adversarial escalation; participant turns were byte-identical. Candidates scored each transcript independently and blind to pairing. A pair passed when the clean twin scored at least 2.0 composite points higher; admission required 18/20 overall and 4/5 in every class. Thresholds and material digests were frozen before scoring, and the pairs remain private.

Table 10: V1 judge-exam administrations and the validity lesson from each. Scores are passed pairs out of 20 unless noted.

Version	Administration, result, and validity lesson
jx1	2026-08-19; all six candidates failed: Gemma-4-31B and MiniMax M3, 15; DeepSeek V4 Pro, 13; Nemotron 3 Ultra, 10; Qwen3.5-397B and the operational Qwen3-235B judge, 9. Sign inversions concentrated in unsafe disclosure: judges often rewarded a mediator's unsupported assertion of a private fact as discovery. The gate blocked panel scoring and exposed a construct flaw rather than a near-margin calibration problem.
jx2	2026-08-19; fresh materials after adding a provenance-and-consent rule; no admission. Gemma scored 16 (4/5 unsafe disclosure), MiniMax 15 (4/5), and the operational judge 7 (0/5). Qwen3.5 failed closed after schema failures on 9 of 40 transcripts. The disclosure-class movement is consistent with repair of the targeted failure mode, not proof of causation because the materials also changed. Sign inversions remained across all four classes.
jx3	2026-08-20; added consent-positive disclosure pairs. GPT-5.6 Sol was admitted at 19 (5/5, 5/5, 4/5, 5/5; exact binomial 95% CI $[0.75, 1.00]$). Grok 4.6 scored 15, Claude Opus 5 scored 12, and Kimi K3 scored 10. This was the first admission across three versions and thirteen candidate administrations. Consent-positive pairs prevented the shortcut of penalizing every private-fact revelation, but a one-member panel could not cover same-lab recusals.

Continued on next page

Version	Administration, result, and validity lesson
Revised protocol	On the same jx3 transcripts, the revised protocol re-scored Opus and Grok and first tested MiniMax. It named vacuous fluency and adversarial reframing and treated caucus disclosure as self-disclosure. Six of Opus's eight failures and three of Grok's five reversed; mean absolute pair-margin shifts were 5.8–6.5 points against a 2.0 decision margin. Grok scored 15, 17, 20, and 17 across variants; the 20 used the non-binding mediator-only yardstick. These comparisons do not separate prompt sensitivity from stochastic rescoring.
Panel close	Sol and Claude Fable 5 were admitted at 19/20 under the frozen rule. Fable joined after the initial jx3 outcome, on same-family-authored materials; thresholds and transcripts were unchanged. Grok was seated by maintainer ruling with a best qualifying score of 17/20 at that decision. Against Sol, McNemar's exact test ($p = 0.219$) did not reject equality or establish equivalence. The ruling cited complementary errors—the three never missed the same pair—and closure of a laboratory-recusal gap.
Test–retest	2026-08-21; with the 40 transcripts, revised protocol, and temperature fixed, Grok's three readings scored 17/20, 19/20, and 15/20; 8/20 pairs changed pass/fail outcome. Median per-pair spread was 3.1 composite points (maximum 6.0), exceeding the 2.0 decision margin. Readings crossed the admission threshold without a wording change.

The exam exposed sign inversions and consent failures but was unstable for fine ranking. Sol's and Fable's 19/20 were single administrations; their repeatability remains unestablished, and Grok's spread does not quantify their uncertainty. A successor should use more items per class and permit an "undetermined" outcome, with prospectively specified tests of semantically invariant perturbations and jury disagreement as a fragility screen [26]. Correctness must be assessed separately from stability. All administrations are reported regardless of direction.

Supplementary Muse Spark candidate. Meta's Muse Spark 1.3 completed the frozen exam on 10 September 2026 through the production judge pathway. It was a first-time candidate added after the earlier outcomes, using the unchanged materials and admission thresholds. Table 11 reports complete coverage and failure of both the overall threshold and the vacuity class floor. This single administration does not establish repeatability or agreement with human judgments. It is separate from the R5 final-scoring instrument and adds no mediator leaderboard result.

Judge	Passing pairs	95% interval	Admission
Muse Spark 1.3 (Meta)	15/20	[0.509, 0.913]	Not admitted

Judge	Vacuity	Partiality	Disclosure	Escalation
Muse Spark 1.3 (Meta)	2/5	5/5	4/5	4/5

Table 11: Judge-admission exam jx3 under its frozen operational evaluator, separate from R5 scoring. Admission requires 18/20 passing pairs overall and 4/5 in every class; a pair passes at an original-minus-sabotaged margin of 2 composite points. An incomplete exam does not establish admission. Scores SHA-256 prefix `bef5aa8d`.

D Model and Run Provenance

Table 12: Human-readable provenance summary. Digest prefixes identify the accompanying machine-readable record; its checksum sidecar preserves full hashes, exact run UUIDs, and complete fields. Hashes identify bytes but do not make private scenario records independently reproducible.

Artifact or field	Compact pin or status	Evidence boundary
Machine-readable provenance manifest	`037d97e68fda`	Full SHA-256 is in `anc/model_provenance_manifest.json.sha256`; the JSON preserves every field summarized here, all exact run UUIDs, and all Sonnet snapshot epochs.
Frozen gate export	`0ab9bb268d37`	Frozen primary-gate aggregate.
July 26 original publication export	`1ad4f6521eab`	Hash-pinned source for historical public run fields and the original mixed-suite analyses.
July 31 corrected public export	`334283df0bf4`	43 publication runs; 134 graph rows; reliability block present; all 103 composite rows carry the coverage contract (94 complete; 9 partial with baseline delta withheld); Sonnet blind 558/558 across 9 runs; August private Qwen runs absent; publication/export provenance only, not paired-estimand input.
Dated aggregate analysis	`c1c53164a144`	Primary and post-freeze aggregate claims; introducing commit `abaff7086941`
Crossed re-score aggregate	`5c5800fbb33c`	36 run–judge pairs at 62/62; historical prompt hash, provider response revision, and job/snapshot UUIDs **NOT RETAINED**.
Sonnet interrupted checkpoint	`6c334a77c2f2`	Preserves the exact preregistered incomplete-coverage checkpoint.
Sonnet complete context readout	`000b450ba372`	9 runs at 62/62; all 36 snapshot UUIDs are retained in the manifest.
Run-registry source commit	`d5193dd80f15`	Prior committed source used only for exact UUIDs and the calibration snapshot.
Scenario corpus (suites 2.0.0 and 2.5.0)	`4ee4ec34003a`	Suite 2.5.0 clones the same corpus hash.
Methodology and calibration	`seeded_25_v1`	Calibration snapshot `e4445e6a-5a69-4b13-86b3-568d87866e4f`
Other historical per-run judge-prompt versions	**UNVERIFIED**	Sonnet is pinned below; other historical inputs remain absent. Current code marker `1.0.0` is not substituted.
Qwen suite-2.5 sensitivity aggregate	`742139a28f04`	8 runs at 62/62; source commit `1a34e8c3abb3`; design registry prefix `dbf63259276b`; 6 historical v1 contracts do not bind participant transport and 2 new v2 contracts do; per-repair attestation not reported by the source endpoint.

D.1 Design amendments

The Gemma-4-31B participant pair was added post-freeze as the registered participant-transport extension (Appendix B, item 10; its own design frozen 2026-07-25, pre-outcome). The frozen design specified one crossed judge (Claude Sonnet 4.5); gpt-oss-120b was added after the freeze but before

Sonnet field	Pinned value	Evidence boundary
Context-withheld coverage	`558/558 COMPLETE`	9 runs; 62/62 each; 36 exact snapshot epochs in the aggregate
Registration	`d98549942138`	2026-07-22T15:15:23Z; withheld-context scoring was prospective; visible reference scores pre-existed
Requested scorer	`claude-sonnet-4-5`	Anthropic; stored labels `Claude45Sonnet` and `Claude45Sonnet::blind`
Provider response revision	**NOT RETAINED** — provider-reported response revision was not retained	the requested alias is not an immutable provider revision
Judge-prompt evidence	suite `1.0.0`; text hash `2b6facca7706`	deployed source-slice hash prefix `bac333355bc5`; prompt text not released
Runtime: visible reference	`558 rows / 9 epochs`	image prefix `8c2510c5be46`; HAI Git prefix `50395648e6af`; nominal commit basis `deployment_tag_and_pulumi_git_head_dirty_tree` (not an in-image source attestation)
Runtime: withheld initial	`390 rows / 18 epochs`	image prefix `a47dc7082c64`; HAI Git prefix `afe5aa5c4f72`
Runtime: withheld resume	`168 rows / 9 epochs`	image prefix `b24797639d1e`; HAI Git prefix `5be524924ddd`

Table 13: Pinned provenance for the complete Sonnet evaluator-context readout. The manifest enumerates all 36 snapshot UUIDs and full hashes; this table reports their three runtime groups. Requested aliases, runtime images, nominal source commits, prompt hashes, and provider-reported revisions remain distinct evidence fields.

Code	Recorded API/model identifier	Serving route and revision evidence
CGLM47	`zai-glm-4.7`	Cerebras; API model id 4.7; immutable serving revision unavailable in retained evidence
F5	`claude-fable-5`	Anthropic; API alias; dated snapshot revision unavailable in retained evidence
G431	`gemma-4-31b`	Cerebras; API model id only; immutable weights/serving revision unavailable in retained evidence
GLM52	`zai-org/GLM-5.2`	Hugging Face router; repository variant 5.2; immutable serving revision unavailable in retained evidence
GPT55	`gpt-5.5-2026-04-23`	OpenAI; dated API id 2026-04-23
H45	`claude-haiku-4-5`	Anthropic; API alias; dated snapshot revision unavailable in retained evidence
K25	`moonshotai/Kimi-K2.5`	Hugging Face router; repository variant K2.5; immutable serving revision unavailable in retained evidence
O48	`claude-opus-4-8`	Anthropic; API alias; dated snapshot revision unavailable in retained evidence
OSS	`gpt-oss-120b`	Cerebras; API model id only; immutable weights/serving revision unavailable in retained evidence
Q235	`Qwen/Qwen3-235B-A22B-Instruct-2507`	Hugging Face router; API/repository variant 2507; immutable serving revision unavailable in retained evidence
S45	`claude-sonnet-4-5`	Anthropic; API alias; dated snapshot revision unavailable in retained evidence
S5	`claude-sonnet-5`	Anthropic; API alias; dated snapshot revision unavailable in retained evidence
STH	`Sao10K/L3-8B-Stheno-v3.2`	Hugging Face router (run registry); release label v3.2; immutable weights/serving revision unavailable in retained evidence

Table 14: Model identifiers copied from the publication export and serving routes checked against the runtime registry. An alias or repository variant is not treated as an immutable provider revision.

Results were pinned, and GLM-5.2 and Kimi-K2.5 were added 2026-07-11–15 with a decision rule for the within-skilled-pair ordering stated before their scores existed (Section 3.4). Every crossed judge — including the post-hoc fifth (Appendix C) — is a strictly offline re-scorer of frozen transcripts, so the amendments cannot affect any generated conversation. GLM-5.2 returned valid scores for 528 of 558 evaluations (94.6%; its missing evaluations concentrate in the longest transcripts, where it exhausts its completion budget on chain-of-thought before emitting scores), so its contrasts are computed over jointly-scored scenario pairs with the joint n reported per contrast; the other crossed judges scored all 558.

D.2 Completion repairs

A run must be complete before we publish it. Any failed scenario row keeps the run out of the public export until we repair it with the original configuration. Repairs are triggered by completion only — failed scenarios or zero-token generations, never by score — and are recorded in the analysis code and internal run records; the public export carries no per-run repair log. One replicate in the Qwen public-context Reflective Listening arm, which the primary E_1 contrast uses, needed two completion-repair passes after provider congestion affected 27 of 62 scenarios. Final coverage was 62/62, and we verified the public-context stamps on the replacement dialogues. The repair, however, crossed serving epochs, and the design cannot separate that epoch effect from generation stochasticity in the replicate.

D.3 Information-study registration chronology

We froze two decision rules on 2026-07-15, before any public-context run existed: E_1 positivity — the lower endpoint of E_1's 90% interval exceeds zero — and E_3 non-inferiority — the lower endpoint of E_3's 90% interval exceeds -4.0 composite points. The margin governs E_3 only. It is not a general materiality threshold, and we specified no equivalence margin for either E_4 estimand. The frozen rules compare the unrounded bootstrap endpoints; the two-decimal values in tables and prose are display values only. The margin's rationale drew on observed full-brief data (about half the frozen full-brief skill increment, and above the observed replicate movement of 3.7 points), but it could not have used any public-context run: the owner signed the rules on 2026-07-15, and the first suite-2.5.0 run launched on 2026-07-17. The dated freeze record is item 9 (Appendix B). On 2026-07-26, after the breadth outcomes were known, we added a separate assay-sensitivity reporting safeguard post hoc: we withhold substantive non-inferiority interpretation where a backbone's E_2 interval includes zero, because the margin could then permit attenuation larger than the observed reference contrast. This later safeguard changes no estimand, threshold, or stored frozen-rule result.

D.4 Bootstrap execution and reporting conventions

Every reported interval on Δ_c comes from a scenario-level cluster percentile bootstrap on the matched pairs $\{(S_c^{\text{med}}(s),\, S_c^{\text{base}}(s))\}_{s \in \mathcal{S}_c^*}$, $n_c = |\mathcal{S}_c^*|$. Read these as scenario-reweighting intervals conditional on the observed generation run or runs. They quantify how the estimator changes when the evaluated scenarios receive different weights; they do not capture variation over newly authored scenarios, fresh language-model generations, or serving implementations. For draws $b = 1, \ldots, B$ with $B = 10{,}000$: (1) resample n_c scenario identifiers from $\mathcal{S}_c^*$ with replacement — whole scenarios are the resampling unit, never individual turns; (2) recompute the *published estimator* of Eq. (2) on the resample. The 90% interval is the percentile interval

$$\text{CI}_{90}(\Delta_c) \;=\; \Big[\, q_{0.05}(\Delta_c^{(1)}, \ldots, \Delta_c^{(B)}),\; q_{0.95}(\Delta_c^{(1)}, \ldots, \Delta_c^{(B)}) \,\Big], \tag{7}$$

with no interpolation: after sorting the B draws, the lower endpoint uses zero-based index $\lfloor 0.05B \rfloor$ and the upper endpoint uses index $\lfloor 0.95B \rfloor - 1$ (for $B = 10{,}000$, the 501st and 9500th order statistics). Because pairs are co-resampled, both arm medians share the resampled scenario set in every draw, so the interval preserves any within-scenario covariance between arms.

The bootstrap RNG is Python's Mersenne Twister (`random.Random`). The operational analysis uses fixed seed 20260706. The crossed-family readout uses the same B, percentile convention, and RNG with seed 20260728, as recorded in its dated artifact. These seeds govern resampling only; LLM token sampling is unseeded and unrelated to them. Within each analysis we re-create the same seeded stream for every reported interval. Draws are coupled where intervals share the same scenario count and traversal convention. Different counts can consume the pseudorandom stream differently, so Monte-Carlo error is not generally common across all intervals, and because seeds are reused, those errors were not designed to be independent. At $B = 10{,}000$ the binomial standard error of an empirical 5%/95% quantile index is about 0.2 percentage points of quantile level, and interval endpoints quoted to two decimals carry that Monte-Carlo resolution. Where a frozen estimand has repeated runs (Section 2.8), we compute the observed per-scenario replicate mean before resampling and then hold it fixed. Only scenario weights vary across draws, so the interval is conditional on the finite set of observed generation runs and contains no run-sampling component.

An earlier implementation resampled replicate identity independently within scenarios, which created hybrid pseudo-runs inconsistent with that scope. The correction changed the emitted intervals, not any estimand, margin, or decision rule.

Three reporting conventions guard against over-reading these intervals.

Discreteness caveat. With $n_c \approx 62$ integer-valued per-scenario scores, the bootstrap distribution of a difference of medians is supported on a small set of observed order-statistic differences. On discrete scores the percentile bootstrap can therefore under- or over-cover in finite samples. Read zero-width or unusually short intervals as discreteness artifacts, not as certainty. The frozen analysis plan prospectively specified smoothed-bootstrap, BCa, and Hodges–Lehmann alternatives (Appendix B).

Within-column contrasts. Two arms in the same participant column share a baseline and a scenario set, so their deltas are strongly positively correlated, and comparing their marginal intervals over-declares ties. We therefore estimate the paired contrast between arms a and a' by co-resampling scenarios and recomputing $\Delta_a^{(b)} - \Delta_{a'}^{(b)}$ per draw, which cancels the shared baseline. Where we show only marginal intervals, we label arm comparisons informal and conservative.

Tie-band decision rule (pre-declared, not a test). The composite HAI Score is the frozen primary estimand; the five components are exploratory. For *ranking* we apply a pre-declared conservative decision rule: two cells or mediators whose 90% intervals overlap are reported as a tie band rather than ordered. This is a ranking convention, not a hypothesis test. Overlap of two intervals is not a claim of no difference, and no p-value is computed or implied.

D.5 Artifact inventory and export chronology

Published results are aggregate and versioned. The public export (`https://mediationbench.com/research/#measurements`) carries, for every run: the recorded participant, mediator, and judge model-identifier strings (including aliases where immutable provider revisions were unavailable), suite and methodology version stamps, run-level scores and deltas, and per-component category

scores; per-role sampling temperatures are recorded once at export level (shared across runs), and per-turn cooperation-curve aggregates with per-point n are present for most but not all archived runs. Sampling is unseeded and disclosed as such.

The export supports inspection of its published run-level aggregates and recomputation of arithmetic derived solely from them, and nothing more: most of this paper's numeric claims — every scenario-paired estimand, scenario-cluster bootstrap interval, and E_1–E_4 value — derive from non-public per-scenario intermediates, because per-scenario scores and deltas are not published under the benchmark's aggregate-only anti-gaming policy. Those quantities are computed by analysis code included alongside the paper (`tools/compute_claims.py` and `tools/crossed_readout.py`; the suite-matched sensitivity additionally uses `tools/qwen_suite25_sensitivity.py`; fixed-seed bootstrap, Section 2.9), with the resulting aggregate outputs archived in-repository as dated JSON artifacts (the post-freeze information-study and participant-transport results are pinned in `analysis/compute_claims_2026-07-28.json`, whose full byte hash is recorded in the generated macro source; the crossed-scorer fixed-transcript aggregates, including the registered post-freeze GPT-5.5 extension reported descriptively in Results, are pinned in `analysis/crossed_readout_2026-07-28.json`, SHA-256 prefix `5c5800fb`; the post-outcome suite-matched Qwen sensitivity is pinned in `analysis/qwen_suite25_sensitivity_2026-08-03.json`, SHA-256 prefix `742139a2`). The aggregate-only Qwen artifact contains eight run means, five estimand values (E_1–E_3 and two E_4 contrasts), and provenance digests, but no scenario rows. The crossed artifact records exact run/judge identifiers, 62/62 coverage, estimator orientation, seeds, and input/generator hashes, but no scenario rows. A separate internal automated adversarial artifact/code review cleared that aggregate for release. The extension was registered before its outcome for outcome-independent reporting; it has one Skilled/full run and no matched Reflective-Listening control. The separately sourced optional operational run score is omitted, not made a dependency of the fixed-transcript claims. Historical crossed-judge prompt hashes, provider-reported model revisions, and rejudge job/snapshot identifiers were not retained; this limits execution-level provenance without changing the hash-pinned aggregate coverage. The frozen gate export contains no suite-2.5.0 run and no rep2 replicate or crossed re-score rows. A later publication export (generated 2026-07-26, archived in-repository with SHA-256 prefix 1ad4f652) adds every generation run used by the original mixed-suite information analysis — including the two suite-2.0.0 Skilled/full runs pooled for Qwen — and the five participant-transport runs, so the generation runs behind E_1–E_4 now have a public, hash-identified record; the crossed-judge re-scores and all per-scenario score matrices remain private. A second immutable public export, generated 2026-07-31 (SHA-256 prefix `334283df`), records the corrected composite rejudge coverage contract and the five-component reliability block: all nine Sonnet context-withheld rows are 62/62, while other incomplete composite rows withhold their baseline deltas. It also includes the registered public GPT-5.5 extension; the two private August Qwen runs are excluded. This corrected export supports export-contract and reliability claims; the dedicated aggregate-only Sonnet artifact remains the source for the paired evaluator-context estimands. The pipeline's built-in validation recomputes run-level means against the published export scores; it does not check the per-scenario medians behind the reported deltas and is silent for runs without a public counterpart — an internal provenance audit, not third-party replication.

The export snapshot this paper cites is archived in-repository as an immutable, hash-identified snapshot (SHA-256 prefix `0ab9bb26`; a DOI at publication), so its public aggregate values remain persistently inspectable after the live export moves on; scenario-level claims retain the reproducibility boundary in Table 9. The judge test–retest sub-study is identified by calibration snapshot `e4445e6a-5a69-4b13-86b3-568d87866e4f`. The 38 reserve fixtures are sequestered and have not been run. The recorded participant identifier strings are pinned: Stheno is a repository label and Haiku an API alias; immutable weights and provider revisions are unavailable in retained evidence.

Their unseeded behavior is stochastic rather than exactly reproducible. The adversarial participant is a Llama-3-8B derivative and is used under the Meta Llama 3 Community License ("Built with Meta Llama 3").

Appendix Tables 12–14 summarize fields pinned by repository evidence, unavailable historical fields, and recoverable release blockers. The accompanying checksum-pinned, machine-readable provenance manifest preserves the full hashes, exact run UUIDs, model fields, and Sonnet snapshot epochs removed from the typeset appendix. In particular, an API alias is not treated as an immutable provider revision, and the current code's prompt-suite marker is not substituted for a historical per-run judge-prompt version.

An internal canary output-leakage scan over all 15,775 messages in the nine frozen runs' live transcripts found zero occurrences of any scenario's canary string in participant or mediator output. This audit detects literal canary leakage in recorded outputs; it neither proves absence of training contamination nor constitutes an externally reproducible check because the scenario-level records and detection mechanics are not public.

D.6 Gate judge corroboration and diagnostics

These details accompany the main judge-reliability results (Section 3.4); they do not change the frozen panel verdict.

Crossed contrasts. On the adversarial column, Skilled@Qwen3-235B minus Skilled@Kimi-K2.5 is +2.3 [0.04, 8.95] under the operational judge (lower bound grazing zero), +7.0 [1.5, 10.0] under the crossed frontier judge, and −2.5 [−6.0, 4.5] under the open-weight judge (sign reversed, interval spanning zero). The cooperative contrast is a tie (+2.1 [−2.5, 4.5]). The frontier judge preserves the ordering that the prospective shared-backbone concern predicted a self-preferring operational judge would inflate (Appendix B). It also preserves the headline contrast nearly unchanged (+35.5 [30.5, 38.0] versus +36.9 [31.9, 41.4]). Of 21 cell×judge estimates, only the cooperative Reflective Listening control under gpt-oss has an interval touching zero (+8.5 [0.0, 13.0]). This family grew after the freeze as judges were added, so the scan is descriptive.

Panel corroboration and GLM sensitivity. On the adversarial column, Kimi's Skilled@Qwen3-235B−Skilled@Kimi-K2.5 contrast is +3.5 [−4.0, 7.0]; GLM's is −5.0 [−17.0, 5.0]. Their cooperative contrasts remain ties (−1.5 [−4.0, 1.0], −19.0 [−30.0, 7.0]). Kimi votes against its own mediator family, so the panel shows no consistent same-family preference. Its re-scores keep every skilled-arm delta and both adversarial skilled-over-control increments interval-backed (+11.0 [3.0, 13.5], +7.5 [1.5, 12.0]). The Sonnet, gpt-oss, and Kimi scorers preserve positive arm-versus-baseline and skilled-over-control contrasts. GLM-5.2 compresses score levels, resolves only some such differences, and reverses the ranked skilled-over-control contrast (−9.0 [−14.0, 4.0], interval spanning zero).

A post-hoc sensitivity could replace GLM-5.2, whose correlations are 0.18–0.28 with other panel members and 0.17 with its own family's GLM-4.7, with that favorable GLM-4.7 probe. This substitution would yield 3–1 and trigger the quirk branch. Neither a reliability floor nor this substitution was prespecified, so the original 2–2 judge-indeterminate verdict stands. The post-hoc GLM-4.7 probe votes with the majority but changes both model version and serving stack; it isolates neither cause and is not a re-vote (Appendix C).

Judge-family generosity. The contrast of Eq. (6), computed against the crossed frontier judge on rounded run-level composites (means of per-scenario composites), is −1 on the adversarial column and +0 on the cooperative column. It shows no operational-judge generosity toward its

own mediator family at the export's one-point resolution. That resolution is coarse relative to a judge-pair gap of up to 4 points; this is descriptive evidence from one judge pair (Section 2.10).

Evaluator-context coverage and shifts. The requested Sonnet 4.5 context-withheld series completed on 2026-07-31 with exact scenario-set coverage across 9 runs: 558 matched run–scenario pairs and 1116 visible/withheld score rows. Registered withheld-minus-visible run-median shifts range over -1.00 to $+3.00$ points; shifts in seven mediated-minus-baseline contrasts range over $+0.50$ to $+3.00$, and shifts in six within-column policy contrasts over -1.50 to $+0.50$. The different serving epochs lack a retained provider response revision. Offline re-scoring cannot alter operational-judge trajectories or stop decisions. The aggregate-only artifact `analysis/sonnet_context_readout_2026-07-31.json` (SHA-256 prefix `000b450b`) discloses exact coverage, snapshot epochs, prompt hashes, and available runtime identities; it publishes no scenario records, score rows, prompts, private briefs, or transcripts.

Test–retest distribution and retention. The pinned calibration grid (snapshot `e4445e6a`) contains all 40×3 passes per component. Cooperative Dimensions is below the tentative-reliability threshold; Resolution Depth is tentative but not reliable. Under the export's three-state confidence label, Resolution Depth, Intent Commitment Symmetry, and Moderator Quality nevertheless qualify as full confidence, while Cooperative Dimensions and Hidden-State Revelations are degraded. 111 of 120 of the raw revelation rating passes occupy the top bucket, whereas the deterministic answer-key F_1 cap pulls the published component toward the floor (Section 3.3). Expected disagreement is tiny: high percent agreement and low α coexist under restricted range. Re-drawing rating passes on the frozen sample yielded Hidden-State Revelations α values of -0.03 and 0.48, with smaller movements on other components. The export retains one execution without an outcome-independent prospective rule; its coefficients must therefore be read with this two-execution range, not as a selected precision gain. Prospective reliability extensions raise n (Appendix B).

E External-Source Pilot

We prepared a 200-case pilot from external, source-labeled material: CaSiNo human role-play negotiations [14] and ContractNLI contract–hypothesis pairs [36]. Only CaSiNo contains conversations. References concern recorded deal acceptance and allocations, and contractual entailment; they are source labels, not newly adjudicated HAI readiness labels. Source-specific reuse conditions are recorded with the sample manifest.

Source	Easy	Medium	Hard	Total
CaSiNo	34	33	33	100
ContractNLI	34	33	33	100
Total	68	66	66	200

Table 15: Prepared cases by source and heuristic difficulty.

Easy, medium, and hard use text and source-annotation heuristics, fixed before judge outcomes, not measured judge difficulty. The manifest records source versions, selected identifiers, selection rules, and hashes. Reference answers and selection metadata stay outside model-visible inputs.

Instrument. Three judge models (Sol, Gemini, and Grok) each read the sample once per reading through a source-specific prompt and a strict JSON parser. This is a new instrument, not the production Formation Judge adapter. Every response must cite verbatim substrings of the turns it can see. A response whose citation does not verify is recorded as failed even when its label matches the source; where that happened, the label-only count follows the strict count in parentheses. Results are reported per source task. There is no pooled accuracy and no R5 composite. The CaSiNo prompt asks only whether a deal was accepted and on what split; the ContractNLI prompt asks only for entailment. Neither asks what the production judges are asked about a negotiation: what became clearer, which terms moved toward agreement, and what remains unresolved. The pilot therefore tests outcome extraction and contract interpretation, not recognition of progress.

First reading (2026-09-05). Table 16. Four defects in our instrument shaped this reading. The CaSiNo inputs included the recorded deal actions, so acceptance was extraction: every judge scored 100 of 100. The prompt described allocations relative to the submitting speaker after the adapter had already named the parties; Sol's 2 allocation misses were exact Party A/B swaps. Both task schemas went to every case, and responses in the wrong shape were classed as technical failures, which opened the circuit breaker: Grok completed 114 cases and never attempted 79. Dispatch ran source by source and stratum by stratum, so every contract Grok attempted was from the easy stratum.

Judge	Deal accepted	Allocation	Entailment	Failed	Pending	Cost (USD)
Sol	100	98	79	0	0	2.39
Gemini	100	100	82	0	0	0.72
Grok	100	100	13	7	79	1.55

Table 16: First reading: strict agreement with the source label, of 100 cases per source. CaSiNo inputs included the recorded deal actions. Cost is the provider-reported or rate-card total per judge.

Second reading (2026-09-06). We withheld every interface action from the CaSiNo inputs, so the judges infer acceptance and split from dialogue alone against the unchanged recorded outcome; sent one task schema per request; recorded wrong-shape responses as judge failures rather than outages; and interleaved dispatch across sources and strata. These changes landed together, so no difference between readings is attributed to one of them. Table 17 reports the result. All three judges attempted all 200 cases; Gemini's 2 failures and Grok's 4 are citation failures.

Judge	Deal accepted	Allocation	Entailment	Failed	Pending	Cost (USD)
Sol	75	69	77	0	0	2.30
Gemini	71	67	82 (84)	2	0	0.75
Grok	69 (71)	64 (66)	77 (78)	4	0	2.19

Table 17: Second reading: deal actions withheld. Strict counts of 100 cases per source; where a citation failure hid a matching label, the label-only count follows in parentheses.

Contracts. Strict agreement with the source label was 77, 82, and 77 of 100 for Sol, Gemini, and Grok (label-only 77, 84, and 78). Aggregate performance is similar, and some labels changed between readings; because the prompt also changed, that does not measure repeatability, and no

ranking is established. Gemini's cost was about a third of Sol's or Grok's. Among the 100 contracts all three judges labeled, 14 received no matching label from any judge, and in 13 of those the judges gave the same label as each other; 9 carry the reference label Contradiction. We report these as shared disagreements with the reference, pending human review of the contracts and of the annotation convention. Further model votes cannot settle which reading the reference should encode.

Negotiations. Table 18 splits the acceptance task by the recorded outcome. With the deal actions withheld, the judges were conservative. Sol missed 22 of 94 recorded deals and claimed a deal for 3 of 6 recorded non-deals; Gemini and Grok were in the same range. All three judges missed the same 22 recorded deals. Their rationales cite splits stated only in part, final proposals with no spoken reply, and quantities that were never spoken: the complete terms lived in the deal form, not in the talk. The 2 recorded non-deals that all three judges called deals each contain a complete spoken agreement followed by a withheld walk-away action, so those judgments are supported by the visible dialogue. When a judge found a deal, its split matched the record in most cases (rightmost column). The defensible finding is that dialogue often fails to establish complete, feasible, mutually confirmed final terms. How many of these deals a person could confirm from the same dialogue is not measured here.

	Recorded deals			Recorded non-deals		
Judge	Found	Missed	Failed	Rejected	Claimed	Exact split
Sol	72	22	0	3	3	66
Gemini	67	27	0	4	2	63
Grok	65	26	3	4	2	60

Table 18: Second reading, CaSiNo acceptance by recorded outcome (94 recorded deals, 6 recorded non-deals). Exact split counts deals found with the recorded allocation.

Scope. These selected public cases cannot estimate deployment error rates or rule out training contamination. Source labels do not establish consent, authorization, or independently verified fulfillment. The full-record protocol below addresses judge preferences with a different instrument.

E.1 Full-record judge preferences (v3)

We applied the simulator Judge's full-evaluation path to the same 100-case CaSiNo sample, using Sol (GPT-5.6-Sol), Gemini 3.8 Flash, Grok 4.6, and Meta's Muse Spark 1.3. Inputs contained the entire shared record, including interface actions in order, with private preferences and source answer keys excluded. Without a mediator, the native mediator-technique rating assesses participant self-management. ContractNLI was not rerun.

Coverage and verification. Table 19 separates full-ledger coverage and cost from scores on the 89 conversations all four judges completed. Sol and Muse exhausted their output budgets on the same case; Grok has 2 interrupted attempts and 8 unattempted cases. Offline replay reproduced requests and parsed results; executable-build differences are recorded with the readout. Each successful case–judge pair contributes one reading.

Judge	Completed	Quality	Self-management	Booked USD
Sol	99/100	8.16	7.53	14.05
Gemini	100/100	8.36	7.34	1.43
Grok	90/100	6.91	5.89	4.78
Meta (Muse)	99/100	8.34	7.10	3.45

Table 19: Full-record CaSiNo judgments. Coverage is out of 100 cases; both rating means use the same 89 completed cases and the native 1–10 scale. Costs use provider-reported amounts or token usage at supplied rates, including failed attempts and $0.22 reserved for Grok's interrupted calls with unknown usage.

Shared ordering, different severity. The six pairwise Spearman correlations on the common sample range from 0.77 to 0.86 for quality and from 0.67 to 0.73 for self-management (average ranks for ties). Grok assigns lower quality than Sol on 83 cases, ties on 6, and assigns higher quality on 0. Gemini completed the entire sample at the lowest recorded cost.

Coarse agreement, finer disagreement. The native progress flag is positive on 87–89 of 89 cases per judge, and all topics are marked resolved on 82–85. All four recommend stopping on every case. Progress is a final-state heuristic, not a measured change, and stopping is not approval to sign. Finer judgments diverge: Sol, Gemini, Grok, and Muse classify the non-participation dimension as cooperative on 47, 82, 60, and 24 cases, respectively; only 14 classifications are unanimous. This field is explicit in the retained answers; other fields can default to "not detected" when omitted.

Interpretation. Human agreement, repeatability, and early stopping accuracy remain unmeasured. R5 uses a different rubric, so comparisons with synthetic results cannot isolate the effect of conversation source.

F Current Synthetic Leaderboard: Exploratory Readout

R5 uses a different instrument from the frozen primary study and the CaSiNo v3 pilot. Each mediator completed one run on the same 62 disputes with Gemma 4 31B parties and a simple goal prompt, as specified by the public scoring method; GPT-5.6 Sol and Gemini 3.8 Flash each scored the saved conversations. Table 20 includes every complete row in the dated public snapshot. Nemotron via Together remained unscored and is excluded from comparisons.

Each score uses Eq. (1)'s component weights (25%, 25%, 20%, 15%, 15%) under the R5 v3 rubric; the means are arithmetic averages across cases. The weights are unchanged, but the rubric and terminal evaluation differ from the frozen study. Each final judge classifies the ending from the transcript independently, with the mediator's stopping declaration withheld: *supported agreement* requires all parties to accept the same specific terms; *supported resolution without agreement* records a substantive resolution without formal agreement; *appropriate no resolution* covers a responsible exit such as impasse, withdrawal, or safety; *premature* means a material issue or unanswered proposal remains without such a justification; and *undetermined* means the transcript does not support a more specific classification.

The judges' displayed means broadly preserve order, with Muse/GLM and DeepSeek/Nemotron reversed. Ending totals differ substantially, so similar aggregate ordering does not establish agreement on stopping. These are marginal counts, not paired case agreement or human accuracy. Single runs

Mediator (provider)	Mean composite Sol	Mean composite Gemini	Premature endings Sol	Premature endings Gemini
DeepSeek V4 Pro (DeepInfra)	49.08	44.02	53	46
GLM 5.3 Flash (OpenRouter)	61.65	70.44	20	2
Gemma 4 31B (Cerebras)	43.88	40.25†	34	34
Grok 4.6 (OpenRouter)	56.11	53.18	22	14
Muse Spark 1.3 (Meta)	63.25	70.33	3	0
Nemotron 3 Ultra (DeepInfra)	47.29	46.31	41	25
Qwen3.8 Flash (OpenRouter)	58.24	61.68	17	5

Table 20: All 7 complete R5 models: scores are 0–100; ending counts are out of 62 per judge. †: Gemini and Gemma share a provider lab. Source: dated public standings, archived 2026-09-07 (SHA-256 prefix `89486ab8`).

provide no stable ranking, and the different rubrics prevent attributing differences from CaSiNo to conversation source.