MediationBench:
Auditing AI Mediation in Synthetic Conflict

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

Working draft — analysis snapshot 2026-08-03. Numbers may change between revisions.

Abstract

We present MediationBench, a synthetic-dialogue harness for dyadic AI mediation. It compares matched mediated and no-mediator conversations across 62 authored scenarios using a 0–100 HAI Score. Synthetic fixtures support controlled information manipulations, authored private facts, and fixed-transcript evaluator audits. The score is not calibrated against human mediation quality or outcomes.

Our strongest result is a prospectively specified fixed-backbone, within-harness descriptive policy contrast: with Qwen and public-context-only information, the Skilled policy exceeds the model-matched Reflective Listening control by $E_1 = +13.17$ [+9.67, +17.55] points (90% scenario-reweighting interval, conditional on observed runs and operational judge). The replicated Qwen core passes its frozen positivity and non-inferiority rules; gpt-oss passes neither. A post-outcome suite-matched sensitivity retains both classifications, but its E_3 interval includes zero, so a positive interaction is not interval-resolved. A registered single-run Gemma participant-transport extension fails the frozen transport criteria. Two offline scorers preserve the negative interaction direction on fixed transcripts; one interval excludes zero. Offline scoring cannot repair trajectories or stop decisions shaped by the operational judge. A preplanned evaluator-context audit is complete; all 22 matched intervals include zero.

Moderator Quality is zero in baselines, so the composite partly encodes treatment; mediated arms bundle asymmetric stopping and extra bandwidth. Intervals reweight observed scenarios rather than generations or human disputes, and withheld scenario-level records preclude external recomputation. Synthetic conversations and synthetic 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

Put two assistant-aligned language models in a simulated dispute, add an AI mediator, and score the outcome: the tested skilled-mediator arms cluster within a point of 54 on a composite 0–100 synthetic-dialogue harness score (the HAI Score, five maintainer-defined rubric components). Under the Stheno-v3.2 high-resistance open-weight participant configuration, the no-mediator cooperation curve never exceeds 7.1% at any observed turn, and the same skilled-mediator arms separate (gate cells 44–48, extension rows reaching 52) over a baseline of 18. These observed ranges show that the tested configurations produce different apparent dynamic ranges; they do not estimate a between-column dispersion interaction or establish which participant property causes the change. The

*Corresponding author: jonathan@hai.io.

pre-declared paired within-column contrasts answer a narrower ordering question: +2.1 [−2.5, 4.5] on the cooperative column (interval includes zero) and +2.3 [0.04, 8.95] on the adversarial column (lower bound only just excludes zero; Section 3.4).

This paper presents MediationBench, the synthetic-conflict testbed that produced the finding. Its measurement infrastructure includes scenario-matched no-mediator comparisons, a minimal Reflective Listening active control, a judge-reliability protocol, and anti-gaming design (cherry-picking, contamination, provenance) disclosed at rationale level in Appendix B. The reported study uses a 62-scenario evaluated/development partition and leaves a 38-scenario sequestered reserve unrun.

The participant model is therefore a design variable rather than an incidental implementation choice: the assistant-aligned pair often resolved disputes unaided while Stheno-v3.2 sustained positional behavior, the two configurations are confounded in capability, disposition, and alignment tuning, and every conclusion is kept conditional on the recorded configuration (Section 2.5).

The study also treats mediator information access as an explicit experimental factor. A model-matched 2×2 crosses mediation policy (Skilled versus the Reflective Listening active control) with information condition (a *full bilateral brief* of both parties' private scenario facts versus *public-context-only* access), yielding four frozen estimands: the public-context skill increment E_1, the full-brief skill increment E_2, their interaction E_3, and the per-policy information contrasts E_4 (defined formally in Section 2.8; results in Section 3.6).

The design separates four descriptive questions that end-state scoring conflates: how mediated arms score relative to their paired baselines, how a skilled policy scores relative to minimal reflective listening (E_1, E_2), how generated scores differ between private-brief and public-context conditions (E_3, E_4), and how an observed ordering behaves under another participant run or fixed-transcript evaluator (the repeat and judge-sensitivity protocols of Section 2.10). The aim is not another model leaderboard; it is an instrument whose sensitivity to each of these choices is measured and reported.

MediationBench delivers two things for builders and evaluators of conversational mediation systems. First, an experimental synthetic-dialogue scoring harness with an observed resolution limit: every mediated frozen gate-cell contrast against its matched baseline has a scenario-reweighting interval above zero, while skilled arms collapse into one tie band per participant column. This bounds what the current composite distinguishes inside the recorded harness; it is not a human-validity result. Second, a versioned benchmark and research artifact that supports controlled information manipulations, matched scenario comparisons, known authored private-information assignments and answer keys, fixed-transcript evaluator audits, repeatable failure analysis, and product regression. Public artifacts are aggregate; private fixtures and scenario-level records remain withheld.

Anti-gaming infrastructure matters for mediation because a system can optimize visible scoring surfaces without improving the interaction being measured. Hidden-state revelation, commitment symmetry, and process trajectories expose distinct rubric signals and candidate failure modes, while the deployment stakes (divorce proceedings, workplace conflicts, public deliberation) make a gameable quality signal worse than none [1].

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, finding that models lag humans particularly in social commonsense reasoning and strategic communication.

Negotiation benchmarks. DealOrNoDeal [7], CraigslistBargain [8], and NegotiationArena [9] score the *negotiating parties'* self-interested outcomes (deal value, win rate); CaSiNo [10] additionally annotates rapport and negotiation strategies, and LLM-stakeholder interactive negotiation [11] 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 [12] and the Habermas Machine [13] showed LLM caucus mediators whose common-ground statements participants preferred over human mediators', and chat-intervention field experiments improved divisive political conversations at scale [14]; these evaluate bespoke systems with human participants. Rehearsal simulates dyadic interpersonal conflict with LLM interlocutors to teach conflict-resolution skill [15]. Recent benchmark work supplies several complementary evaluation units. 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 — a human-annotated reference this benchmark lacks; it evaluates mediator decisions rather than running complete mediated and no-mediator conversations on matched authored scenarios [16]. ProMediate executes proactive agents in multi-party, multi-issue negotiation settings and measures consensus change, intervention timing, and social intelligence against no-agent and generic-agent conditions [17]. SoCRATES constructs 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 [18]; its criticism that existing testbeds rely on a few expert-authored domains applies to our authored suite, whose 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 — the nearest precedent for our information estimands and participant axis; our E_4 differs in comparing information conditions within a model-matched policy pair against a matched no-mediator baseline [19]. 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 [20] and exhibit systematic biases [21]; we operationalize this literature's recommendations as benchmark infrastructure (a fixed operational judge distinct from all participants and mediators, a crossed second judge, and test–retest reliability per component).

Synthetic social-simulation validity. Model-on-model interaction can expose prompt, policy, information, and evaluator sensitivity under controlled conditions, but shared linguistic priors can also create compatibility that does not transport to people or even to another participant simulator. 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 [22]. We therefore treat participant-family transport and human criterion calibration as separate validity gates, not as consequences of a high synthetic score.

Our work differs from these neighboring testbeds in measurement emphasis, not categorical priority. The methodological novelty we claim is the pairing discipline: per-scenario matched mediated/no-mediator pairs on an identical authored partition, with a within-scenario co-resampled bootstrap that preserves cross-arm covariance. A shared baseline cancels in a contrast of two deltas only when both reuse that same baseline; it does not cancel inside an individual mediated-minus-baseline delta. Scores are therefore reported as matched descriptive deltas rather than absolute conversation quality; generation remains unseeded, so each delta also contains stochastic-trajectory differences, which prospectively specified replication addresses but does not eliminate. The scope is deliberately *dyadic* — a narrowing relative to ProMediate's and SoCRATES' multi-party settings, accepted to keep the matched-pair estimand clean and the participant axis interpretable. Two further emphases: process is measured alongside outcomes (per-turn cooperation dynamics and hidden-state revelation) under formal anti-gaming mechanisms, and participant-model selection — including the Stheno-v3.2 high-resistance open-weight configuration — is treated as a first-class validity variable rather than an implementation detail.

2 Methodology

2.1 Design overview

The frozen gate design is a fixed-harness factorial matrix: two participant configurations × four arms, the complete 62-scenario evaluated/development partition, one replicate per cell (496 conversations; one retained Reflective Listening replicate is additional and reported separately as a noise anchor), under a single operational judge and one suite version (2.0.0). Factors:

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; see caption) 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). The Gemma-4-31B participant pair was added post-freeze as the registered participant-transport extension (prereg 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 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.

The Reflective Listening arm is a *minimal-intervention active control*: it acknowledges and reflects participant messages but does not apply the structured mediation policy. It is therefore an active intervention comparator, not an inert-presence placebo. The Skilled arm applies that structured policy: beyond acknowledgment, it probes for undisclosed constraints and interests, names tradeoffs explicitly, and pushes toward specific reciprocal commitments. The verbatim policy prompt is withheld under the same rationale-level disclosure policy as the operational judge prompts (Appendix B); this behavioral description is the level at which the policy is published. Keeping Skilled@Qwen3-235B in the roster while Qwen3-235B-A22B-Instruct is the operational judge doubles as a judge–mediator generosity probe under the crossed judge (Section 2.10).

2.2 Scenario design

For a benchmark, the fixture design is the instrument, so we state the authoring protocol. The inventory contains 100 authored fixtures: the 62-scenario evaluated/development partition and the 38-scenario sequestered, unrun reserve. All were authored for the benchmark by the maintainers, with LLM writing assistance under a written authoring guide and human curation before inclusion; none derives from a real dispute, transcript, or personal data (Ethics statement). 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 B). The evaluated fixtures also informed benchmark development; the reserve exists precisely so a future locked epoch can test generalization beyond them.

2.3 The need for baselines

Absolute scores are uninterpretable in isolation. Every scenario is run *without* a mediator under the same seeded protocol to establish a matched baseline; mediated scores are then reported as deltas against the matched baseline (same suite version, participant model, judge model, and scenario set). The Moderator Quality component contributes 0 for baseline runs, so the maximum baseline score is 85. A mediated arm scoring at the baseline shows no positive composite-score difference in this comparison. A negative delta is flagged, not suppressed; it may reflect over-intervention, bias, escalation, or generation-path variation and therefore warrants transcript-level inspection rather than a single causal diagnosis.

2.4 Run protocol

For each scenario: (1) the scenario fixture turns are persisted verbatim as a shared seed transcript; (2) participants continue the conversation until each has 25 public turns; (3) in mediated arms the mediator may intervene after the seed and may stop early only on a model-reported agreement event or a judge `should_stop`, while the baseline arm has no stop event by design; (4) the operational judge scores the full transcript with the standard HAI Score formula; per-round judge evaluations are persisted for the cooperation curve in all arms (measurement-only in the baseline arm). Three consequences of (3): 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. Runs must be complete to publish: any failed scenario row excludes the run from the public export until repaired with the original configuration. Repairs are completion-triggered — failed scenarios or zero-token generations, never score — and are documented in the analysis code and internal run records; the public export does not carry a per-run repair log. One replicate in the Qwen public-context Reflective Listening arm used by the primary E_1 contrast required two completion-repair passes after provider congestion affected 27 of 62 scenarios. Final coverage was 62/62 and the public-context stamps were verified on the replacement dialogues, but this repair crossed serving epochs; the current design cannot separate that epoch effect from generation stochasticity in the replicate.

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, and its own 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 in the prospectively specified, frozen 2×2 information study supplies the shared scenario context and the public dialogue as it unfolds, but not either party's private facts.

The full bilateral brief is a privileged-information provisioning condition; preparation through intake or caucus is only an imperfect analogy for it. It is not an observed or simulated interview: no elicitation occurs, and the researcher-authored private facts are supplied verbatim. Accordingly, the contrast should not be read as an estimate of the value of real-world interviewing or case preparation. Because the active control and skilled policy receive the same information within each condition, their score increment compares policies at fixed information. Information was fixed across the original gate arms; 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 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 D) crosses policy and information within each backbone.

Judges (the operational judge and every offline panel judge) receive the same reconstructed judge-facing transcript and content under the harness: the scenario title and description, both parties' backstories, archetypes, and stated goals, and the full transcript — but never the hidden facts and never the Hidden-State Revelations answer key, whose overlap cap is applied mechanically in code after judging. Two implications are drawn where they bite: 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. Model inference, guessing, or hallucination can nevertheless 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 D).

2.5 Participant configurations as a validity variable

Participants in benchmark conversations are simulated by LLMs configured as disputants. Participant selection 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, producing high baselines and compressed mediator-arm differences. Stheno-v3.2 sustained more positional behavior and yielded lower no-mediator baselines and a wider observed point-score range in this stress test.

These observations do not validate either configuration as a model of human behavior, nor do they identify why the configurations differ. Model capability, conflict disposition, and alignment tuning change together and are therefore confounded. “Open-weight” describes access to the model artifacts; it is not itself an explanation for the observed behavior and 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 requires a future factorial participant study; the present comparison supports only configuration-conditional claims.

External evidence makes the alignment-tuning component of this confound independently plausible: ablating the learned safety-refusal direction [23] within a fixed open-weight backbone moves instruction-tuned models’ closed-ended General Social Survey responses closer to matched human response distributions, and recovers suppressed mind attribution — approaching the human panel average for non-human animals while rising to or above it for chatbots, technology, and natural entities — with theory-of-mind and general-reasoning accuracy left intact [22]. That result concerns closed-ended survey responses from small open-weight models, not interactive conflict behavior, so it validates neither participant configuration here. It does support treating alignment tuning as a distinct axis of the confound, and its within-backbone ablation design is a template for the factorial study above: safety-ablated and unablated variants of one participant backbone would separate an alignment-tuning manipulation from backbone and capability, which the present cross-model comparison cannot.

2.6 The HAI Score: what we measure

The HAI Score (0–100) is a maintainer-defined composite that scores five specified features of generated conversations inside this harness; Table 2 lists its components and weights, which Eq. (1) combines. Human criterion validity is unestablished.

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; they are 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.

The composite is a maintainer-specified settlement-progression and information-surfacing construct, 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.

Because no mediator exists in baseline runs, Moderator Quality is structurally set to zero there; mediated-versus-baseline deltas therefore mix four shared outcome components with one treatment-specific component. We report the component-level results and return to this limitation in Section 4.

Three component mechanics determine the reported values and are stated here as specification. Resolution Depth is a mean over detected topics, so every topic that is surfaced but not closed lowers it. 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, so sparse detection cannot read as fully cooperative. Intent Commitment Symmetry defaults to 0.5 when a conversation contains no detected commitments, so commitment-free transcripts read as symmetric rather than asymmetric. Hidden revelations are compared with seeded expected facts using deterministic text-overlap precision, recall, and F_1; the F_1 acts only as a cap on the revelation component. The public export can also report per-component inter-rater reliability — Cohen's κ, Krippendorff's α, sample size, rater kinds, and calibration time — when those ratings are available.

2.6.1 Hidden-state revelation: construct and measured floor

The framework tracks information visibility through three states: *hidden* (known only to one party; strategic advantage, blocks collaborative problem-solving), *hinted* (vaguely referenced; signals openness while preserving optionality), and *revealed* (fully disclosed; enables mutual problem-solving). At the construct level, a consequential constraint remains *hidden* while it is known only to one party, becomes *hinted* when that party alludes to an unspecified constraint, and becomes *revealed* when the party makes it explicit enough for the conversation to address it. Breakthrough moments in conflict resolution frequently occur when hidden interests behind stated positions become explicit [24, 25], so a mediator that helps place a constraint on the table can advance collaborative problem-solving when the affected party voluntarily consents and confidentiality is protected — safeguards this score does not measure. The information condition (Section 2.4) scopes that interpretation: under the full bilateral brief, revelation measures whether known authored private-information assignments — supplied verbatim in the mediator's packet — are surfaced in conversation, not unaided discovery, and it cannot by itself distinguish elicitation skill from use of the brief. In the present snapshot this component sits near floor in every arm (Section 3.3) — the construct is retained because it is the benchmark's most direct process claim, and its floor behavior is itself diagnostic: the judge assigns high raw revelation ratings in the test–retest sample while the deterministic overlap cap drives the scored component to the floor (Section 3.4).

AI participants also generate distinct internal-thought and public-message streams, allowing a belief–behavior divergence signal (the gap between hidden intent and stated position). It is not summed into the composite and is not populated for the runs in this snapshot; it remains future instrumentation (Appendix D).

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), combined into the per-scenario HAI Score

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

with the weights fixed constants of the scoring harness (Section 2.6). Eq. (1) is the continuous per-scenario composite 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, which are persisted as rounded integers, so every crossed-judge delta and interval endpoint lies on a half-integer grid while operational-judge values are continuous (a granularity difference flagged where 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 is public, complete, contains *no* failed scenario, and agrees with the mediated run on all four matching keys: suite version, participant model, judge model, and the exact scenario set $\mathcal{S}$. Pooled cross-model baseline statistics are never used. Post-freeze repeat runs are reported and analyzed separately rather than automatically pooled. The sole disclosed replicate pooling in this paper is the information study, which takes a per-scenario *mean* of the arm's pinned replicates before the scenario median, as specified in Section 2.8.

Estimand. Let $\mathcal{S}_c^* \subseteq \mathcal{S}$ be the scenarios with a defined score on *both* arms. Writing $S_c^{\text{med}}(s)$ and $S_c^{\text{base}}(s)$ 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,

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

Note Δ_c is a difference of medians, not a median (or mean) of paired differences; the alternatives disagree on skewed data, and only the former reproduces the reported interval-backed deltas. Two aggregations circulate and are kept 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), computed over the jointly-completed subset. The two do not coincide even when both arms are complete on the identical scenario set (on the headline cell the export delta is +30 while $\Delta_{\text{med}} = +36.9$), which is why Table 3 reports both. Every Δ_c is reported per cell: no delta is pooled across participant models, arms, judges, or suite versions. It 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 and crosses policy $p \in \{\text{Sk}, \text{RL}\}$ (Skilled; Reflective Listening) with information condition $i \in \{\text{full}, \text{pub}\}$ (full bilateral brief; public-context-only). For arm (p, i), let $M_{p,i}$ denote the median, over the scenarios jointly completed in all four arms, of the per-scenario score — where an arm has more than one replicate run, the per-scenario score is the mean of its replicates. The four frozen estimands are

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

Intervals come from a four-arm scenario-co-resampled cluster bootstrap: scenarios jointly completed in all four arms are resampled once per draw (preserving cross-arm correlation), each arm's observed per-scenario replicate mean is held fixed, and every draw recomputes the same functional as the point estimate. The interval is therefore conditional on the recorded generations; it does not estimate run-to-run variation. An earlier implementation independently resampled replicate identity within scenarios, creating hybrid pseudo-runs inconsistent with that scope. The correction changed the emitted intervals, not an estimand, margin, or decision rule.

Two decision rules were frozen 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 no equivalence margin was specified for either E_4 estimand. The frozen rule's comparisons use the unrounded bootstrap endpoints; values rounded to two decimals in tables and prose are display values only. The margin's rationale referenced observed full-brief data (about half the frozen full-brief skill increment, and above the observed replicate movement of 3.7 points); the input it could not have used is any public-context run — the rules were owner-signed 2026-07-15 and the first suite-2.5.0 run launched 2026-07-17. The dated freeze record is prereg item 9 (Appendix D). A separate assay-sensitivity reporting safeguard was added post hoc on 2026-07-26, after the breadth outcomes were known: substantive non-inferiority interpretation is withheld where a backbone's E_2 interval includes zero, since the margin could then permit attenuation larger than the observed reference contrast. This later safeguard changes neither estimand, threshold, nor stored frozen-rule result.

2.9 Uncertainty: scenario-reweighting bootstrap

All reported intervals on Δ_c use a scenario-level cluster percentile bootstrap on the matched pairs $\{(S_c^{\text{med}}(s),\, S_c^{\text{base}}(s))\}_{s\in\mathcal{S}_c^*}$, $n_c = |\mathcal{S}_c^*|$. They are best read 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, not 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

$$\mathrm{CI}_{90}(\Delta_c) \;=\; \Big[\, q_{0.05}\big(\Delta_c^{(1)}, \ldots, \Delta_c^{(B)}\big),\; q_{0.95}\big(\Delta_c^{(1)}, \ldots, \Delta_c^{(B)}\big) \Big], \tag{4}$$

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, the two arm medians share the resampled scenario set in every draw, so the interval preserves whatever within-scenario covariance exists 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 and are unrelated to LLM token sampling, which is unseeded. Within each analysis the same seeded stream is re-created 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. Seed reuse means 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), the observed per-scenario replicate mean is computed before resampling and then held fixed. Only scenario weights vary across draws; the resulting interval conditions on the finite set of observed generation runs and contains no run-sampling component.

Three reporting conventions guard against over-reading these intervals; they are one deliberate apparatus, stated once here.

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, so on discrete scores the percentile bootstrap can under- or over-cover in finite samples; zero-width or unusually short intervals should be read as discreteness artifacts rather than certainty. Smoothed-bootstrap, BCa, and Hodges–Lehmann alternatives were prospectively specified in the frozen analysis plan (Appendix D).

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

2.10 Judge reliability protocol

The judge is a measurement instrument with noise, and we measure it rather than assume it away. Reliability is computed 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 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

$$\alpha \; = \; 1 - \frac{D_o}{D_e}, \tag{5}$$

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. Because Cohen's κ is a two-rater statistic, no pairwise κ is reported 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; ordinal-distance variants were prospectively specified in the frozen analysis plan. These coefficients characterize per-component bucket agreement; no reliability coefficient is computed 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, which is the denominator of every inter-judge statistic below; it 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

$$\kappa \;=\; \frac{p_o - p_e}{1 - p_e}, \qquad p_e \;=\; \sum_{k=1}^{5} p_k^{(1)}\, p_k^{(2)}, \tag{6}$$

and nominal Krippendorff's α on the same paired codes; Section 3.4 reports the composite-level agreement and run-level reproduction this block yields. The same offline protocol was subsequently extended to the full four-judge panel (gpt-oss-120b, GLM-5.2, Kimi-K2.5; Table 1 caption), whose 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 exhibit systematic biases [20, 21]; our operational judge is distinct from every participant model, and one mediated arm shares the operational judge's backbone, enabling 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. Because the visible reference scores predate registration and the two conditions span different serving epochs, 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 judge $j \in \{Q, S\}$ (operational, crossed) assigns to the transcripts of mediator backbone $m \in \{Q, K\}$ — 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

$$\mathrm{SP} \;=\; \big(\widetilde{S}_{Q,Q} - \widetilde{S}_{S,Q}\big) \;-\; \big(\widetilde{S}_{Q,K} - \widetilde{S}_{S,K}\big) \tag{7}$$

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. $\mathrm{SP} > 0$ 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. Judge-level uncertainty is unavailable from a single judge pair; a scenario-cluster bootstrap can show SP's scenario-reweighting variation conditional on the observed transcripts but cannot supply judge-population uncertainty. We report SP descriptively alongside the separate multi-judge sensitivity panel.

3 Results

The gate aggregates are anchored to the frozen public-export snapshot (SHA-256 prefix `0ab9bb26`); post-freeze analyses are identified where reported. Scenario-reweighting intervals were computed from scenario-level records with the committed analysis pipeline (`tools/compute_claims.py`; scenario-cluster percentile bootstrap, B=10,000, fixed seed). The aggregate public export alone does not contain the records needed to reproduce those intervals. The complete 8-cell matrix appears in the frozen snapshot with 62/62 evaluations per cell; the evaluated development partition itself is not public.

Overview. Three results organize this section. First, the replicated Qwen public-context-only contrast: the Skilled policy exceeds the model-matched control by $E_1 = +13.17$ [+9.67, +17.55] and passes both frozen rules; gpt-oss passes neither, and a post-outcome suite-matched sensitivity retains the classifications while its E_3 interval includes zero (Section 3.6). Second, the registered Gemma participant-transport extension fails the frozen transport criteria, and two offline judge families preserve the negative interaction direction on fixed transcripts. Third, the instrument audits — a four-judge crossed panel and a complete evaluator-context audit — leave the fine skilled-pair ordering judge-indeterminate while arm-versus-control contrasts stay positive under the three higher-correlating crossed judges (Section 3.4).

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.

Every frozen gate-cell interval in Table 3 excludes zero — a descriptive result under the frozen estimand and harness; post-freeze extensions are labeled exploratory where they appear. The benchmark's central structural observation is that **the tested participant configurations produce different observed point-score ranges in this snapshot.** Under the cooperative participant the baseline is already 36 and every skilled mediator collapses into a narrow band (54–55), ceasing to separate from one another; under the adversarial participant the baseline falls to 18 and the same mediators span a wide range. The magnitude components of this column structure

— baseline levels and delta sizes — appear 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 participant configurations jointly differ 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 published mediated gate cell scores above its matched baseline, with larger absolute differences under the adversarial participant (+30 and +26 versus +18 and +18). Those aggregate gains coexist with negative scenario-level score differences under the flagged-not-suppressed rule of Section 2.3: between 0–8 of 62 scenarios per gate cell score *below* their matched generated baseline — 0 scenario scores for Skilled@Qwen3-235B under Stheno-v3.2 (L3-8B), and 8 for the Reflective Listening control under Claude Haiku 4.5. In selected baseline re-runs, repeating the identical no-mediator configuration moved 56 (adversarial) and 57 (cooperative) of 62 scenario scores in one direction or the other (Section 3.5). This demonstrates material generation stochasticity but is neither a random sample of repeats nor a bound on how much of any negative-delta count is policy-associated. The active minimal-intervention control sets a more demanding comparator than no mediator alone: it recovers +22 of the +30 achieved by the strongest gate-cell Skilled policy.

The control's gain splits into two parts, one of which is the scoring rubric rather than change in the four shared components. Baseline runs score Moderator Quality as 0 by construction (Section 2.3) and the judge awards the Reflective Listening control 90.8 on that component, so about 13.6 points of the control's *mean* composite gain are rubric arithmetic mechanically available to any mediated arm. The remainder is a descriptive arm difference in the four shared components, including Cooperative Dimensions (Section 3.3); it cannot be attributed specifically to interleaved reflection because the arms also bundle stopping, conversational bandwidth, and generation stochasticity. The analysis contrast used here is the Skilled-policy increment above this active control — the gate skilled arms lead it by +4 to +8 points on the adversarial column (+4 to +12 including the single-replicate extension rows of Section 3.5). It is distinct from the platform's raw median-of-N leaderboard rank, which does not bind any paper row (Appendix B). The prospectively specified co-resampled contrast (Section 2.9) makes this interval-backed for one arm: the Skilled@Qwen3-235B increment over the control is +7.8 [2.8, 17.7] under the operational judge and 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.

The post-hoc sensitivity registered after component results were known is reported for all nine gate rows in Table 4. 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.

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 the Δ itself, and it is data-dependent (Moderator Quality saturates in this snapshot, Section 3.3). The two control runs differ by one point (40, 41); across four targeted post-freeze repeat cells the maximum observed run-level movement is 3.7 composite points (Section 3.5). Those selected repeats do not bound run-to-run variation and remain short of the prospectively specified paired-by-seed design (Appendix D). This active-control decomposition is therefore a decomposition of the mean gain, not a significance claim about any mediator pair.

Participant	Arm	Run	Published Δ	Shared-four Δ	Mod. Qual. contribution (share)	MQ-excluded Δ
Claude Haiku 4.5 (coop.)	No mediator (baseline)	`110ce648`	+0.0	+0.0	+0.0 (—)	+0.0
	Reflective Listening	`82958a25`	+13.0	-1.5	+14.3 (112.1%)	-1.8
	Skilled@Qwen3-235B	`74e861d6`	+18.0	+2.8	+14.8 (84.0%)	+3.3
	Skilled@Kimi-K2.5	`e051b7f9`	+18.0	+2.7	+15.0 (84.7%)	+3.2
Stheno-v3.2 (L3-8B) (adv.)	No mediator (baseline)	`20012f5a`	+0.0	+0.0	+0.0 (—)	+0.0
	Reflective Listening	`d69386c3`	+22.0	+8.5	+13.6 (61.7%)	+10.0
	Reflective Listening (replicate)	`77c0a672`	+23.0	+9.7	+13.6 (58.5%)	+11.4
	Skilled@Qwen3-235B	`0928d6af`	+30.0	+15.7	+14.2 (47.5%)	+18.4
	Skilled@Kimi-K2.5	`2068ccb2`	+26.0	+11.5	+14.4 (55.7%)	+13.5

Table 4: Post-hoc Moderator Quality sensitivity for all nine frozen gate rows. The four-component value removes the 15% mediator-only component and renormalizes the remaining weights. These are run-level aggregate sensitivities, not replacements for the frozen composite or algebraic decompositions of the difference-of-medians estimand.

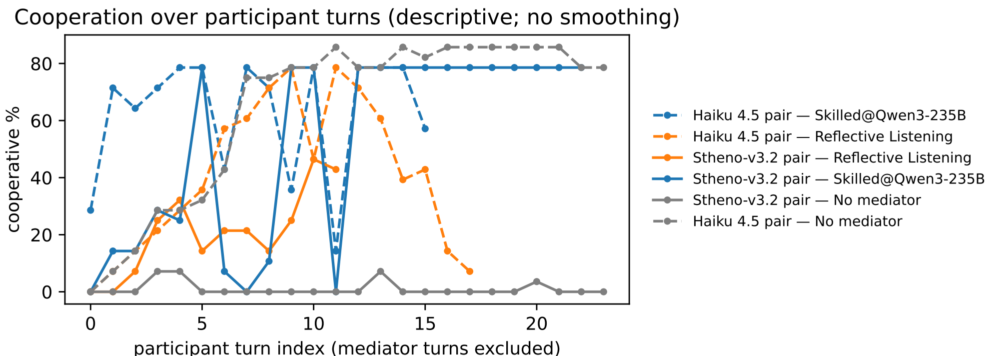
3.2 Cooperation over participant turns

The curves (Figure 1) carry the intuition behind the column structure. The adversarial no-mediator baseline never exceeds 7.1% median cooperation at any of its 24 observed turns — the disputants simply keep fighting — while the cooperative-column baseline resolves largely unaided. The descriptive areas under the per-turn median curves separate accordingly: 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, they do not share a common turn support across arms, and mediated-arm curves are survivorship-conditioned per the caption.

Under this protocol, mediated gate conversations ran shorter than baseline conversations. Baselines run the full protocol (53.2 and 58.4 total transcript turns per scenario on the cooperative and adversarial columns), while every mediated gate arm ends 46%–65% earlier (18.6–28.7 turns per scenario). The registered administrative-censoring caveat applies here: baseline conversations are censored at the protocol cap, so time-to-agreement is estimable only within mediated arms (Appendix D, item 3). Two readings follow. First, the survivorship conditioning above is quantitatively large, not a technicality. Second, 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 using 35% fewer turns. Turn totals are published per run in the export; this is a descriptive observation, not a ranked outcome.

3.3 Per-component breakdown (exploratory)

Table 5 reports every gate row's per-component category scores. Four component-level facts are stable across the snapshot and frame how the composite should be read. First, discrimination between arms is carried almost entirely by 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). Second, Moderator Quality *saturates*: it scores 90.79–99.98 across every mediated arm in the snapshot, gate and extension alike — including the Reflective Listening control at 90.8 — so 15% of the composite functions as a mediated-arm indicator rather than a skill measure. This is direct evidence for labeling the composite weights maintainer-asserted, and it motivated the completed Moderator Quality-excluded, renormalized sensitivity reported in Table 4. That analysis was registered only after the component results were known, was not in the frozen plan, and never replaces the frozen primary estimand. Third, Intent Commitment Symmetry moves *against* treatment in three of the four adversarial mediated rows and in the cooperative Skilled@Kimi-K2.5 cell (Table 5). Fourth, Hidden-State Revelations, the construct's nominal



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
n at risk: Haiku 4.5 pair — Skilled@Qwen3-235B	62	59	46	33	23	12	7	5	4	3	2	1	1	1	1	1							
n at risk: Haiku 4.5 pair — Reflective Listening	62	60	59	54	45	37	29	22	16	12	8	6	3	2	2	2	1	1					
n at risk: Stheno-v3.2 pair — Reflective Listening	62	61	59	58	46	35	25	17	12	10	4	2											
n at risk: Stheno-v3.2 pair — Skilled@Qwen3-235B	62	59	46	37	24	15	7	5	4	3	1	1	1	1	1	1	1	1	1	1	1	1	1

Figure 1: Cooperation over participant turns for the six roster gate cells (per participant column: no-mediator baseline, Reflective Listening control, and strongest gate skilled arm; column by linestyle, arm by colour). Participant-turn axis; mediator turns excluded. Points are per-turn medians across scenarios; descriptive, no smoothing. The table beneath gives the per-turn number of active conversations (n-at-risk) for the mediated arms; baseline arms have no stop event, so their n stays at 62. Per-turn values are conditional on conversations still active at that turn: mediated runs stop early on agreement or judge stop while the baseline has no stop event, so late-turn mediated composition shifts toward unresolved scenarios.

primary progress indicator, stays near floor (8–17) in every arm including baselines. The test–retest distribution resolves the source: 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 drives the published component toward the floor; this mechanical disagreement does not establish criterion-valid detection, and component-level conclusions there are unsafe.

3.4 Judge reliability

Crossed re-judge (identical transcripts). Two judges from other model families re-scored all 8 gate cells plus the Reflective Listening replicate (nine runs, 62 transcripts each) offline, on the identical frozen transcripts the operational judge scored. The crossed frontier judge (Claude Sonnet 4.5) reproduces the operational judge's run-level composites within 0–4 points on every cell and preserves the full mediator ordering — including Skilled@Qwen3-235B above Skilled@Kimi-K2.5 under the adversarial participant, the ordering that the prospectively documented shared-backbone concern (Appendix D) predicted a self-preferring judge would inflate. The second offline judge (gpt-oss-120b, open-weight) scores at a systematically lower level and preserves the control-versus-skilled separation, but *not* the ordering within the skilled pair: under it the adversarial-column Skilled@Kimi-K2.5 delta edges above Skilled@Qwen3-235B (Table 6). The prospectively specified paired contrast quantifies this: on the adversarial column, Skilled@Qwen3-235B minus Skilled@Kimi-K2.5 is +2.3 [0.04, 8.95] under the operational judge (the lower bound grazes zero) and +7.0 [1.5, 10.0] under the crossed frontier judge, but −2.5 [−6.0, 4.5] — sign-reversed, interval spanning zero — under the open-weight judge; on the cooperative column the pair is a plain tie (+2.1 [−2.5, 4.5]). 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; correlation does not establish absolute score agreement. Median

Participant pair	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
	Reflective Listening	72.1	39.1	12.4	28.8	95.5
	Skilled@Qwen3-235B	79.1	45.8	14.0	32.8	99.0
	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
	Reflective Listening	45.3	29.1	12.2	36.2	90.8
	Reflective Listening (replicate)	49.8	29.8	10.2	38.2	90.8
	Skilled@Qwen3-235B	64.2	41.5	10.3	34.5	94.5
	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).

deltas under all three judges appear in Table 6; the headline contrast survives the crossed judge nearly unchanged (+35.5 [30.5, 38.0] vs. +36.9 [31.9, 41.4]). Of the 21 cell×judge estimates — a family extended post-freeze as panel judges were added, so this is a descriptive scan, not a frozen inferential claim — the single interval touching zero is the weakest treatment under the harshest judge (cooperative-participant Reflective Listening control under gpt-oss, +8.5 [0.0, 13.0]).

Four-judge panel adjudication of the ordering (post-freeze). To adjudicate the sign reversal we extended the crossed re-judge to a four-judge cross-family panel — Claude Sonnet 4.5, gpt-oss-120b, GLM-5.2, and Kimi-K2.5 — over the same nine frozen runs, with the decision rule stated before the two new judges produced a single score: three or more of four crossed judges with the adversarial Skilled@Qwen3-235B−Skilled@Kimi-K2.5 contrast positive would report the reversal as one judge's quirk; a 2–2 split would report the within-pair ordering as judge-indeterminate. The rule was stated after the first two crossed judges' scores existed but before the two added judges produced any score, and it specified no 1–3 branch; only those two outcomes were defined. The panel split 2–2: positive under Claude Sonnet 4.5 (+7.0 [1.5, 10.0]) and Kimi-K2.5 (+3.5 [−4.0, 7.0]); negative under gpt-oss-120b (−2.5 [−6.0, 4.5]) and GLM-5.2 (−5.0 [−17.0, 5.0]). Only the frontier judge's interval excludes zero, and the cooperative column stays a four-judge tie (−1.5 [−4.0, 1.0], −19.0 [−30.0, 7.0]). The prospectively specified verdict is therefore *judge-indeterminate*.

Two panel facts sharpen the read. First, Kimi-K2.5 is the mirror image of the shared-backbone concern — a judge scoring its own family's mediated arm — and it votes *against* itself, placing Skilled@Qwen3-235B ahead; the panel shows no consistent same-family preference pattern. Second, GLM-5.2 is a measured outlier. Its scenario-level Spearman rank correlation with every other judge is 0.18–0.28, against 0.65–0.80 among the operational, frontier, gpt-oss, and Kimi judges (all-judge joint set, n=528). Kimi-K2.5's re-scores keep every skilled-arm delta and the adversarial skilled-over-control increments interval-backed (+11.0 [3.0, 13.5], +7.5 [1.5, 12.0]); GLM-5.2, by contrast, compresses levels, resolves only a subset of the arm-versus-baseline and skilled-over-control differences, and sign-reverses the ranked skilled-over-control quantity itself (−9.0 [−14.0, 4.0], interval spanning zero). That reversal comes from a low-correlating scorer with incomplete coverage — its missing evaluations concentrate in the longest, arm-correlated transcripts, and the panel rule counts unweighted signs — but disagreement among unvalidated LLM judges does not by itself establish which scorer is wrong. Under any reliability floor that would exclude a judge correlating at 0.18–0.28

with every panel member — and at 0.17 with its own family's GLM-4.7 (below) — the counterfactual vote is 3–1 and the rule's quirk branch fires; no floor was prespecified, so the registered verdict stands as 2–2. The panel accordingly refines the earlier summary: Skilled@Qwen3-235B versus Skilled@Kimi-K2.5 is a tie band whose within-band ordering is judge-indeterminate by pre-stated rule; arm-versus-baseline and skilled-over-control contrasts remain positive under the Sonnet, gpt-oss, and Kimi scorers, while the low-correlation GLM-5.2 scorer compresses or reverses them.

A post-hoc fifth-judge probe of the GLM-5.2 anomaly (GLM-4.7 on a different serving stack) votes with the majority but cannot separate model-version from serving-stack sensitivity; it is a sensitivity probe, not a re-vote, and is reported in Appendix C. The 2–2 verdict above was prospectively specified over the original panel and stands.

The prospectively specified judge-by-arm generosity contrast (Eq. (7)), computed on the export's run-level composite scores (means of per-scenario composites, stored as rounded integers) against the crossed frontier judge, is −1 on the adversarial column and +0 on the cooperative column: at the export's one-point integer resolution — coarse against a judge-pair run gap of up to 4 points — the operational judge shows no generosity toward its own family's mediated arm on either column. This is a descriptive read on one judge pair, per Section 2.10.

		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). The separately frozen re-score under the requested Sonnet 4.5 alias asks whether an offline evaluator's access to the parties' private briefs changes its reading of otherwise identical transcripts. The context-withheld series completed on 2026-07-31 at exact scenario-set coverage for all 9 runs (558 matched run–scenario pairs and 1116 score rows across the visible and withheld conditions). The registered withheld-minus-visible run-median shifts range from −1.00 to + 3.00 points. Shifts in the seven mediated-minus-baseline contrasts range from +0.50 to + 3.00, and shifts in the six within-column policy contrasts range from −1.50 to + 0.50. All 22 90% scenario-resampling intervals include zero. Thus this epoch-separated fixed-transcript readout does not show a directionally resolved context effect on run or policy-contrast medians for this scorer. It is not evidence of equivalence: no equivalence margin was registered, the intervals permit nonzero shifts, and visible and withheld scores were produced in different serving epochs whose provider response revision was not retained. Because the re-score is offline, it also cannot alter trajectories or stop decisions made by the operational judge. The aggregate-only artifact is `analysis/sonnet_context_readout_2026-07-31.json` (SHA-256 prefix `000b450b`); it discloses exact coverage, snapshot epochs, prompt hashes, and available runtime identities without publishing scenario records, score rows, prompts, private briefs, or transcripts.

Intra-judge test–retest. The operational judge re-scored 40 stratified transcripts (five per gate run) at K=3 passes each (calibration snapshot `e4445e6a`); the pinned rating grid is complete — 40×3 passes per component. The 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 Hidden-State Revelations value requires its distribution to interpret. 111 of 120 of its rating passes sit in the *top* bucket: the judge's raw revelation ratings are high even as the answer-key F_1 cap pulls the published component score to floor (Section 3.3). Expected disagreement is therefore tiny and α is variance-starved. Percent agreement is high, and the low α is strongly affected by restricted range in this sample; it does not exclude instability specific to the component. That fragility is directly observable: the sub-study executed twice against this frozen sample (rating passes re-drawn) returned Hidden-State Revelations α of -0.03 and 0.48 across executions, with smaller movements on the other components. The export identifies one execution as the calibration snapshot of record, but no outcome-independent rule for retaining that execution was prospectively specified; its coefficients therefore must be read together with the disclosed two-execution Hidden-State Revelations range, not as a selected precision gain. Per-component α point estimates at n=40 under restricted range should be read with that draw sensitivity in mind (the prospectively specified reliability extensions, Appendix D, raise n). The export schema attaches a three-state confidence label to each published coefficient; under its criteria Resolution Depth, Intent Commitment Symmetry, and Moderator Quality qualify as full confidence while Cooperative Dimensions and Hidden-State Revelations are flagged degraded. The two components that carry arm discrimination (Cooperative Dimensions, Resolution Depth) sit at 0.54 and 0.74. Cooperative Dimensions is degraded and below the tentative-reliability threshold; Resolution Depth clears that tentative threshold but not the reliable threshold. The tie-band rule (Section 2.9) does not absorb it: its intervals reweight scenarios and contain no judge-rescoring variance. Judge noise is measured by this sub-study and the panel, and the paper's own data show the distinction — the within-skilled-pair contrast excludes zero under the rule's intervals, is judge-indeterminate under the panel, and flips under re-generation (Section 3.5).

Judge-swap generation rows — separate conversations under a different operational judge, in which judge influence is confounded with generation stochasticity — are superseded by the crossed re-judge above and reported in Appendix C.

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

Beyond the frozen matrix, 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 now 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 every interval 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 marginal-interval-based, conservative read; the paired contrasts of Section 2.9 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 worth recording because it 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 with no run-sampling interval and only a few points of separation; their scenario-reweighting intervals remain conditional on the one observed generation. The observation is a hypothesis-generating one that the prospectively specified replicated design (Appendix D) can test, namely whether within-harness mediation-policy performance is distinct from general model strength.

End-to-end replicates (post-freeze). Four cells were re-run end-to-end after the freeze — 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. Run-level composites move at most 3.7 points: cooperative baseline 36 $\rightarrow$ 37.9, adversarial baseline 18 $\rightarrow$ 15.5, Skilled@Kimi-K2.5 44 $\rightarrow$ 46.2, Skilled@Qwen3-235B 48 $\rightarrow$ 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), which is not an equivalence test, and 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 demonstrate generation stochasticity: the identical no-mediator configuration moved 57 (cooperative) and 56 (adversarial) of 62 scenario scores in one direction or the other. Because these are 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 replicate-level observation bears directly on the panel verdict of Section 3.4: the within-skilled-pair point ordering *flips* across replicates (frozen: Skilled@Qwen3-235B 48 > Skilled@Kimi-K2.5 44; runs: Skilled@Kimi-K2.5 46.2 > Skilled@Qwen3-235B 44.3) while both repeated mediated-minus-baseline score-difference point estimates remain positive. In every frozen gate aggregate, repeat runs are analyzed 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 do not alter any frozen gate value. Together with the four-judge panel, these selected repeats 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)

The prospectively specified information-provisioning manipulation (Appendix D) ran as a model-matched 2×2 on the adversarial column: Reflective Listening and Skilled policies, both bound to the same Qwen backbone (that of Skilled@Qwen3-235B), crossed with the full-bilateral-brief versus public-context-only conditions defined in 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, which 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 directional sensitivity a reader can check from Table 7. The Qwen matrix has two runs per cell, every run 62/62, with estimands, analysis code, and the −4.0-point non-inferiority margin 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 has point scores similar to the original control (39/41 vs. 40), supporting comparability on Qwen rather than backbone invariance.

The strongest result is the prospectively specified, contemporaneous public-context-only contrast: under Qwen, the Skilled policy exceeds the model-matched Reflective Listening active control by $E_1 = +13.17$ [+9.67, +17.55] composite points. This is a 90% scenario-co-resampled interval conditional on the observed Qwen runs and the operational judge Qwen3-235B-A22B-Instruct; it is neither a run-sampling interval nor an estimate of performance in human disputes. 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] against the prospectively frozen −4.0-point non-inferiority margin. The registered Qwen verdict is therefore 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 latter interval includes zero and does not establish equivalence.

Mediator backbone	Replicates	E_1: public skill	E_2: full-brief skill	E_3: interaction	$E_{4,\text{reflection}}$	$E_{4,\text{skilled}}$
Qwen	2/2/2/2	+13.17 [+9.67, +17.55]	+7.27 [+1.31, +10.52]	+5.90 [+1.34, +13.81]	−5.89 [−12.89, −2.52]	+0.01 [−4.15, +4.28]
Kimi	2/1/1/1	+9.66 [+2.32, +16.21]	+1.68 [−2.06, +5.85]	+7.98 [−0.32, +15.47]	−6.45 [−11.99, +0.60]	+1.53 [−1.64, +5.75]
gpt-oss	1/1/1/1	−1.02 [−4.51, +5.01]	+4.74 [+2.19, +15.55]	−5.75 [−16.75, +0.38]	+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.

One descriptive point-estimate 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], +1.53 [−1.64, +5.75], and −1.64 [−4.59, +3.38]). Each of these three $E_{4,\text{skilled}}$ intervals includes zero. The Qwen value is additionally suite-dependent: the suite-matched sensitivity below moves it from +0.01 [−4.15, +4.28] to −2.66 [−6.40, +2.04]. This does not establish equivalence: no $E_{4,\text{skilled}}$ equivalence margin was specified, the intervals remain compatible with nontrivial shifts, and the full-brief Skilled-policy arms are historical. Read against the already-frozen ±4.0-point band as a labelled post-hoc diagnostic (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]) exceeds 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 — instructively, it is the replicated core whose interval is widest. The companion reflective-policy contrast also carries 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, which bounds how uniformly the "clusters near zero" reading can be extended 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. E_2, E_3, and the Skilled-policy information contrast combine those newer cells with protocol-matched frozen suite-2.0.0 full-brief Skilled-policy arms. Run date and suite stamp therefore differ even though the arm protocol was held fixed, so historical drift or other version-linked differences can affect those three estimates. The design gives the cleanest reading to the public-context-only Skilled-versus-reflective contrast, not to a universal claim that skill replaces information.

As a post-outcome, suite-matched sensitivity, 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`); here "post-outcome" means that the sensitivity was chosen after observing 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 (see the model-provenance appendix). 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 the interaction's positive direction is not interval-resolved under the suite match; 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 since 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 and does not establish equivalence. This post-outcome sensitivity does not replace the original mixed-suite primary analysis and is neither a replication nor a wholly contemporaneous factorial: its intervals reweight the fixed scenarios conditional on the observed generations and do not estimate run-to-run uncertainty. The aggregate-only artifact is pinned by short hash `742139a2`.

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). The mostly single-run Kimi matrix crosses both frozen numerical thresholds. Its E_2 interval ($+1.68$ [-2.06, $+5.85$]) includes zero, however, 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. gpt-oss crosses neither frozen threshold: 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, so Kimi is a threshold-crossing but interpretation-limited breadth check, and gpt-oss is a failed generalization; there is no backbone-universal positive skill increment or interaction result. This single-run failure identifies a policy–backbone configuration that did not generalize; it does not establish that gpt-oss is generally weak, because 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). The frozen 2×2 was replayed on a second participant configuration — Gemma-4-31B participants (prereg item 10) — with the Qwen mediator backbone, both policy prompts, both information conditions, the operational judge, and the scenario partition held 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 transported: 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]). The public-context policy contrast did not transport: $E_1 = +5.70$ [−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. 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; it establishes that the Stheno public-context finding did not transport to this participant configuration under the frozen rules. 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). 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 additionally stores two scoped exploratory run × judge pairs not analyzed in the paper's Results, 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. First, baseline clearance is preserved under both offline scorers: 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-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]), preserving the positive full-brief reference contrast under the later assay-sensitivity safeguard. And the negative interaction is preserved in 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 provide 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. Applying that margin descriptively would therefore 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. 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 provides 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. (An earlier same-day emission used the median of within-scenario paired differences rather than the registered difference of arm medians and computed $E_2 - E_1$ rather than the registered $E_1 - E_2$. Adversarial re-verification caught both implementation errors before any snapshot pinned them. Only the implementation and its emitted values were corrected: the estimand, margin, and decision rules did not change. The registered-estimand computation above is the release-ready aggregate readout.)

4 Limitations and Threats to Validity

We state these plainly; several are design choices whose costs we accept knowingly, and several map to a prospectively specified remedy in Appendix D.

1. **Replication is uneven across claims.** The reported intervals quantify scenario reweighting for a fixed configuration, not run-to-run variance. Post-freeze, four cells (both baselines, both adversarial-column skilled cells) were re-run end-to-end: run-level composites moved at most 3.7 points, the frozen-minus-repeat baseline intervals include zero, and the within-skilled-pair point ordering flipped (Section 3.5). These are maximum observed movements in four selected repeat cells, not bounds on run-to-run variation; an interval including zero is not evidence of equivalence. This remains far from the prospectively specified paired-by-seed design. The Qwen information-context matrix has two replicates per cell; the Kimi and gpt-oss breadth

matrices are mostly single-run conditional probes, and gpt-oss fails the prospectively specified generalization reading.

2. **The scenario suite is fixed and authored.** The bootstrap distribution reflects reweighting of these scenarios, not random sampling from a defined population of human disputes. Its intervals cannot establish prevalence or effect size in real conflicts. The 62-scenario evaluated development partition and scenario-level records are not public; the aggregate public export alone cannot reproduce the reported scenario-reweighting intervals.

3. **Purposively selected participant configurations, not a participant population.** The two gate configurations plus one registered exploratory extension show that participant configuration changes what this harness measures; they support no pooled population-level claim. The extension covers the participant axis through the exploratory Gemma participant-configuration extension of the information factorial (Appendix D, item 10; results in Section 3.6) — single-run-per-arm conditional breadth that holds the mediator and evaluator apparatus fixed within its cells, removing some package confounding without isolating a participant property. Both frozen rules failed on that column.

4. **Capability, disposition, and alignment tuning are confounded** between participant configurations. Claude Haiku 4.5 and Stheno-v3.2 (L3-8B) differ jointly in participant backbone, capability, training and alignment, disposition, serving provider, and stochastic generation. The results cannot attribute the observed point-score-range difference to any participant property. The same-model persona-swap deconfound remains prospective; the safety-ablation deconfound of Section 2.5 is a candidate design, not a registered item.

5. **The active control's backbone was unpinned in the frozen matrix.** The Reflective Listening moderator has no model binding, so its calls fell to the harness default (`claude-sonnet-4-5`; disclosed in the export's `mediator_model` field), while skilled arms ran their named backbones. The frozen skilled-over-control contrast therefore decomposes mediator *packages*, not policy at fixed backbone (Section 2.4); the prospectively specified model-matched 2×2 removes this confound for Qwen, not for all backbones. Relatedly, mediator prompts window to the last 20 messages while the system prompt (which carries the full bilateral brief in full-brief arms) is re-sent every call — an asymmetry disclosed as part of the prospectively specified public-context-only estimand.

6. **The primary information factorial is not wholly contemporaneous.** E_2, E_3, and the Skilled-policy information contrast combine contemporaneous suite-2.5.0 cells with protocol-matched frozen suite-2.0.0 full-brief Skilled-policy arms, so run date and suite stamp differ for those three estimates; the suite-matched sensitivity removes the mismatch but is post-outcome and remains noncontemporaneous. The full argument appears once, in Section 3.6.

7. **The mediated treatment is a bundle.** Mediated arms may stop on agreement or judge advice, baseline conversations cannot stop, and mediator turns do not consume the participants' turn budget. Deltas therefore combine intervention content with stopping policy and extra conversational bandwidth. A shared global budget and symmetric stop rule, as used to control interaction resources in ProMediate [17], are needed to isolate these factors.

8. **Single operational judge; LLM-as-judge noise.** All conversations are steered and scored by one judge model. Test–retest and crossed-judge sub-studies measure — but do not eliminate — instrument noise; judge self-preference is probed descriptively only. The registered

cross-family re-score of the suite-2.5.0 rows (Section 3.6) provides narrower evidence than an end-to-end judge swap: two offline scorers on fixed transcripts corroborate the negative transport interaction's direction, while magnitude and uncertainty remain evaluator-dependent. Neither scorer can correct trajectories or stop decisions already influenced by the operational judge.

9. **Judge-swap baselines are separate conversations.** Baseline generations under different operational judges differ in both instrument and trajectory; they are not re-scores and cannot isolate judge influence.

10. **Component weights are maintainer-asserted**, not derived, and Moderator Quality is structurally zero for no-mediator runs but saturates for mediated arms (including the Reflective Listening control). The headline delta therefore mixes a common four-component outcome with a treatment-specific component. The completed Moderator Quality-excluded, renormalized four-component analysis is post hoc and secondary (Table 4); it does not replace or algebraically decompose the frozen difference-of-medians estimand.

11. **Arm non-blindability.** The judge observes mediator turns in mediated arms; scoring cannot be blinded to arm.

12. **Survivorship conditioning in per-turn curves.** Mediated conversations stop early on agreement or judge stop; the baseline has no stop event. Per-turn values are conditional on still-active conversations and every curve point carries its n.

13. **Token sampling is unseeded.** Conversations are stochastic; `sampling: unseeded` is exported rather than implying a determinism that does not exist. Paired-by-seed replication is prospectively specified.

14. **Bootstrap endpoint discreteness.** With ∼62 paired deltas, percentile endpoints land on observed order statistics, so on discrete scores the percentile bootstrap of a difference of medians can under- or over-cover in finite samples; smoothed-bootstrap, BCa, and Hodges–Lehmann alternatives are prospectively specified.

15. **Delta-level attrition.** Δ_c is estimated over scenarios completed on both arms; differential scenario failure induces selection into the pair set, and the matched-baseline rule rejects an entire baseline run on any single failed scenario. (All gate cells published to date are 62/62 complete, so this is currently latent.)

16. **Reliability scope.** Test–retest and crossed-judge coefficients are computed on per-component 1–5 buckets; no reliability coefficient is computed for the continuous composite or for Δ_c itself. Unlike ProMediate's human metric validation and SoCRATES' expert-alignment study [17, 18], this version has no human calibration of scenario realism, mediator quality, judge scores, or downstream human outcomes. Human criterion validity for mediation quality is therefore unestablished.

17. **Preparation is an analogy, not an observed practice.** Neither information condition (Section 2.4) observes an interview, intake, or confidential human caucus. The 2×2 tests information availability inside this harness, not the field performance of mediator preparation, interviewing, or mediation. Full briefs supply both parties' private facts without modeling elicitation, confidentiality, consent, or permission to disclose. One measured provisioning nuance qualifies the public-context-only label: scenario `focus_areas` reach the mediator in

every arm, including public-context-only, and 3 of the 100 authored fixtures have a focus area sharing enough content words with an expected revelation to act as a directional hint — not a verbatim answer key — so the public-context condition is not perfectly information-free.

18. **Simulated, narrow disputes.** Participants are language models with configured hidden states, not human subjects. The suite is dyadic, English-language, text-only, and excludes nonverbal behavior, institutional power, cultural interpretation, and multi-party coalition dynamics. External validity to human conflict is unestablished.

5 Conclusion

We presented MediationBench, an experimental synthetic-dialogue scoring harness that measures how generated dyadic conversations score under an AI mediation policy, using paired no-mediator baselines, an active Reflective Listening control, a process-level composite score, and a judge-reliability protocol, under anti-gaming mechanisms disclosed at rationale level. It complements mediator-output, multi-party negotiation, and cross-domain adaptation testbeds by auditing the measurement consequences of participants, information, and judges. In a fixed, frozen harness, the tested participant configurations produce different observed point-score ranges: against the cooperative configuration, skilled-mediator point scores occupy a narrow band, while the tested higher-resistance configuration produces a wider spread. The configurations remain confounded in capability, disposition, and alignment (Section 4), so the result does not identify a causal participant property. Every published mediated gate cell exceeds its matched baseline, and the minimal Reflective Listening control itself scores above that baseline, while the Skilled-policy increment must be estimated above that active control. Part of the control's apparent gain is mechanically assigned by the mediator-only rubric component.

A four-judge cross-family panel re-scoring identical transcripts records the fine ordering within the skilled tie band as judge-indeterminate under its rule pre-stated before the two added judges scored; arm-versus-control contrasts remain positive under the three mutually higher-correlating crossed judges, while the low-correlation GLM-5.2 scorer compresses or reverses them. Four targeted end-to-end repeats show a maximum observed run-level movement of 3.7 points but do not bound run-to-run variation. A preplanned Sonnet evaluator-context-withheld re-score reached complete coverage across all nine registered series, and all 22 matched 90% intervals include zero, leaving the context effect directionally unresolved for that scorer. A secondary observation — that mediator ordering under higher-resistance participants does not obviously track general model strength — is hypothesis-generating from single-replicate data, not a result. In the separate information experiment, the strongest result is the contemporaneous Qwen public-context-only contrast: the Skilled policy exceeds the model-matched Reflective Listening active control by $E_1 = +13.17$ [+9.67, +17.55] composite points (90% scenario-reweighting interval conditional on the observed runs and operational judge). Mostly single-run Kimi crosses both frozen numerical thresholds, but a later, explicitly post-hoc assay-sensitivity safeguard withholds its E_3 interpretation because E_2 includes zero; gpt-oss crosses neither frozen threshold. The interpretable decision-rule outcome therefore does not generalize across the three tested backbones. A post-outcome, suite-matched Qwen sensitivity leaves the frozen classifications unchanged while widening the interaction interval to include zero. One descriptive pattern does recur: the Skilled-policy information contrast clusters near zero in point estimate across all three tested backbones — a hypothesis-generating pattern, not equivalence, suite-dependent for the replicated backbone (the suite-matched sensitivity moves Qwen's estimate to -2.66 [-6.40, $+2.04$]), for the reasons stated in Section 3.6. It motivates a contemporaneous, single-suite, paired-seed study with a prospectively specified equivalence margin and does not imply the full bilateral brief can be removed. Full bilateral briefing versus public-context-only access is a controlled analogue of preparation and intake, not an observed interview or evidence about human mediation efficacy.

The failed participant-transport result is itself a useful assay outcome: the registered single-run Gemma extension failed the frozen transport criteria for the Qwen public-context finding, and two offline scorers preserved the negative E_3 direction on fixed transcripts.

The practical implication for evaluation is concrete: score synthetic mediation against both a matched no-mediator baseline and an active minimal-intervention control; treat mediator information access as an explicit experimental factor; stress participant configurations; and cross-score frozen transcripts with cross-family judges. The benchmark's claims are bounded — uneven replication, two confounded gate participant configurations plus one exploratory participant-transport extension, an uncalibrated simulated population, historical-arm reuse in parts of the factorial, and a measured-noisy judge — and the prospectively specified plan states what evidence would license stronger ones. 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 data-pipeline invariants can serve as hard product regression gates. Fixed-transcript LLM scoring is instead a controlled statistical regression test that needs frozen tolerances and calibrated false-alarm rates; unseeded end-to-end generated scores remain advisory under the same constraint.

Priority next steps are human expert review of transcripts and judgments for criterion calibration; paired-seed, contemporaneous single-suite replication; a same-model persona swap to separate disposition from backbone and alignment; and broader judge and participant-family coverage.

Ethics Statement

This study involves no human subjects: all disputants are simulated by language models with configured scenario roles and hidden states, and all transcripts are machine-generated. The author did not seek institutional review because the study uses no human subjects, human-derived transcripts, or personal data. Scenario fixtures were authored for the 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 [26], and engagement-optimized ranking systematically amplifies out-group animosity [27]. 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/results/`) 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.

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 re-scores.
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 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 10–15 separate fields pinned by repository evidence from unavailable historical fields and recoverable release blockers. 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.

DRAFT NOTE: Before submission: deposit the release inventory in Zenodo and record its DOI; resolve every recoverable field still classified as unresolved in Table 10. 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] 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.

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

[9] 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.

[10] 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.

[11] 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".

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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/.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

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

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

[26] 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.

[27] 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.

[28] Allan Gibbard. Manipulation of voting schemes: A general result. *Econometrica*, 41(4):587–601, 1973.

[29] Mark A. Satterthwaite. Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2):187–217, 1975.

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

A Model and Run Provenance

Table 10: Repository-pinned provenance and unresolved release fields. Hashes identify bytes; they do not make private scenario records independently reproducible.

Artifact or field	Pinned value	Repository evidence
Frozen gate export	`0ab9bb268d37c919da483d92db54f6670fcc1407cf2a7ef1f439f7563bb280c0`	`docs/research/exports/latest_2026-07-06T0913Z_sha0ab9bb26.json`
July 26 original publication export	`1ad4f6521eab582545391c0577b2c1262900538697c82cd1501055df505f9d52`	`docs/research/exports/latest_2026-07-26T0722Z_sha1ad4f652.json`
July 31 corrected public export	`334283df0bf4a25f9751fb06d68b762d7e29a04bfdb809d8787365bdcd3f12a8`	`docs/research/exports/latest_2026-07-31T0402Z_sha334283df.json`; generated 2026-07-31T04:02:19.183346195Z; 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; checksum in `docs/research/exports/latest_2026-07-31T0402Z_sha334283df.json.sha256`
Dated aggregate analysis	`c1c53164a1441186e7369b8a1a3592a3dcf55021510322a68f0ef1d0c7e266b9`	`docs/research/analysis/compute_claims_2026-07-28.json`
Crossed re-score aggregate	`5c5800fbb33cf8e997b86b0a4540e3069966b2c9454cb874ea65fcf9fb7458c0`	`docs/research/analysis/crossed_readout_2026-07-28.json`; 36 pairs at 62/62; release-ready; checksum in `docs/research/analysis/crossed_readout_2026-07-28.json.sha256`
Sonnet interrupted checkpoint	`6c334a77c2f2d6a107216af98e62d99d3b3791c36de288ee8212e7235fbaddd6`	`docs/research/analysis/sonnet_context_withheld_coverage_2026-07-28.json`; checksum in `docs/research/analysis/sonnet_context_withheld_coverage_2026-07-28.json.sha256`

Continued on next page

Artifact or field	Pinned value	Repository evidence
Sonnet complete context readout	000b450ba372cf3e0fdf7409 9c79f8cbc8a7d95cb4690fb6 d947fef1954ea3ca	docs/research/analysis/sonnet_context_readout_2026-07-31.json; 9 runs at 62/62; checksum in docs/research/analysis/sonnet_context_readout_2026-07-31.json.sha256
Analysis artifact introducing commit	abaff7086941f2d4a0da129e 38b7019d88933db3	unique introducing commit verified against repository history
Run-registry source commit	d5193dd80f150bd4ad50f042 1247182003cd7dd8	prior committed docs/research/paper/tools/compute_claims.py used only for exact UUIDs and the calibration snapshot
Scenario corpus (suites 2.0.0 and 2.5.0)	4ee4ec34003ae7ca6d9a2f88 983de56e05ee2898a3c6bbbc c8d44aafc39b5e07	api/src/benchmarking/scenarios.rs; suite 2.5.0 clones the hash in api/migrations/293_blind_reference_mediators.sql
Methodology version	seeded_25_v1	publication export top-level field
Calibration snapshot	e4445e6a-5a69-4b13-86b3-568d87866e4f	docs/research/paper/tools/compute_claims.py
Current code prompt-suite marker	1.0.0	current api/src/benchmarking/run_inputs.rs; not historical run evidence
Other historical per-run judge-prompt versions	**UNVERIFIED**	the Sonnet series is pinned below; other historical run inputs remain absent from checked artifacts
Crossed re-score historical execution provenance	36 run/judge pairs pinned; historical prompt hash, provider response-model revision, and job/snapshot UUIDs **NOT RETAINED**	exact run/judge identities are in docs/research/analysis/crossed_readout_2026-07-28.json; the aggregate records these historical limitations explicitly
Qwen suite-2.5 sensitivity aggregate	742139a28f041055008b015f 56fd02b9af2a22ff7e76b69a 7cbfe8af9c35ec21	docs/research/analysis/qwen_suite25_sensitivity_2026-08-03.json; 8 runs at 62/62; source commit 1a34e8c3abb3b0a868ae3937 e98c7d7984dc6a51; design registry dbf63259276b60a61bf1e5c5 fda7518e2509ef39587cb4b8 e655d3b6ca643fa7; 6 historical v1 contracts do not bind participant transport and 2 new v2 contracts do; per-repair attestation not reported by the source endpoint; checksum in docs/research/analysis/qwen_suite25_sensitivity_2026-08-03.json.sha256

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	`d985499421384bb32c3a247832e23ff3e864c452`	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-reported response model	**NOT RETAINED** — provider-reported response revision was not retained	response model was not retained; the requested alias is not an immutable provider revision
Judge-prompt evidence	suite `1.0.0`; text hash `2b6facca77062f07386e60cfc5d0cf874f237e507bd13ee36eb684b8b10a9b6a`	deployed source-slice hash `bac333355bc51eb359c2ae66d4d7bd4609ef071a34f5266b1a9a8e8be18b9c85`; prompt text not released
Runtime: visible reference	`558 rows / 9 epochs`	image `8c2510c5be46218284895c38f40cf42c4e5a97d46611e0a76721d376158c2b24`; HAI Git `50395648e6af20f0d4f6f78a5b53ae2461ad71b6`; 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 `a47dc7082c648569b9f17391fc88c87b0581fba12b9ba2a8c74ba41f44e43048`; HAI Git `afe5aa5c4f72590bf993cd9f15f43d354d166350`
Runtime: withheld resume	`168 rows / 9 epochs`	image `b24797639d1e253c5fa59eb2e2e48bd1366a8eba4eaf36b730388a0f1d9d6cae`; HAI Git `5be524924ddd3166e2ff4677f14e298062811c88`

Table 11: Pinned provenance for the complete Sonnet evaluator-context readout. The aggregate enumerates all 36 snapshot UUIDs; this compact table reports their three runtime groups. Requested aliases, runtime images, nominal source commits, prompt hashes, and provider-reported revisions are distinct evidence fields.

Code	Recorded API/model identifier	Configured route (code evidence)	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 12: 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.

Run UUID	Suite	Participant	Mediator	Judge
`0928d6af-a940-4a55-8ba5-20bf21445c87`	`2.0.0`	STH	Q235	Q235
`546966c9-f3e0-4d9b-948e-d0f81a6eae29`	`2.0.0`	STH	Q235	Q235
`77c0a672-ad9d-49fe-8537-195cd4ead15e`	`2.0.0`	STH	S45	Q235
`d69386c3-2786-40d3-9e77-9853e9931af4`	`2.0.0`	STH	S45	Q235
`00d8ce02-c7ae-4a91-a84e-8ab059e49441`	`2.0.0`	STH	S5	Q235
`579b3609-7270-48b5-b2ec-08ec46d7233e`	`2.0.0`	STH	OSS	Q235
`2068ccb2-818c-4b74-bb75-5993dd4e5503`	`2.0.0`	STH	K25	Q235
`d05fcf37-7270-40d1-a077-0ee9de582ec7`	`2.0.0`	STH	K25	Q235
`20012f5a-63c8-47e2-9426-00239ed1f177`	`2.0.0`	STH	—	Q235
`8f356f48-0e1c-41be-bc07-25c06a8d9cc5`	`2.0.0`	STH	GLM52	Q235
`74e861d6-f7f9-4bdc-8761-14da705f62f8`	`2.0.0`	H45	Q235	Q235
`725cbbd6-e9ee-4ed9-9999-299fb321bac0`	`2.0.0`	H45	F5	Q235
`97af1740-1756-4ad1-967e-c68049de1422`	`2.0.0`	H45	O48	Q235

Table 13: Exact UUIDs and recorded model fields for runs addressed by the dated analysis artifact (part 1).

Run UUID	Suite	Participant	Mediator	Judge
82958a25-51dd-41a4-a240-cb9ee831ca35	2.0.0	H45	S45	Q235
50693d02-d298-4097-87a9-5aadfbd628b2	2.0.0	H45	GPT55	Q235
356d0940-abb2-4e67-a6fd-b93bcf194a0c	2.0.0	H45	OSS	Q235
e051b7f9-4964-40fe-b2f0-2fd6ae7f628b	2.0.0	H45	K25	Q235
110ce648-edb0-40f5-bcc4-de5371ccb134	2.0.0	H45	—	Q235
a28950e9-30b1-46ad-80ce-0850abd86c23	2.0.0	H45	GLM52	Q235
0e3e0f47-e0cf-4a82-be0e-42bd2432e8d2	2.5.0	STH	Q235	Q235
77e6b874-3163-4955-ba13-aa6d0f4031d4	2.5.0	STH	Q235	Q235
83752dc7-8edf-4a3d-8a5e-aceeb6e29cb0	2.5.0	STH	Q235	Q235
986c4f8c-bbf4-4c11-9f09-11b3c7a39d43	2.5.0	STH	Q235	Q235
df31ef89-7280-4ce6-b625-db84c31687f0	2.5.0	STH	Q235	Q235
f0b1d261-9f3d-40b6-95df-f1b8b6245149	2.5.0	STH	Q235	Q235
5f378266-2439-40ac-aedc-24aceb79ef3e	2.5.0	STH	OSS	Q235

Table 14: Exact UUIDs and recorded model fields for runs addressed by the dated analysis artifact (part 2).

Run UUID	Suite	Participant	Mediator	Judge
cd46badc-69cc-4bec-a22a-60604567c7fb	2.5.0	STH	OSS	Q235
d82f5d2a-1fd3-4c5b-b011-a28661c7e4d5	2.5.0	STH	OSS	Q235
1b711261-7c3d-47c3-866b-2f7dc415a490	2.5.0	STH	K25	Q235
726730c8-6791-43ad-ad3b-02e50bd7465c	2.5.0	STH	K25	Q235
83a48744-cb7b-4794-a5a2-93c7c930603d	2.5.0	STH	K25	Q235
0f77a820-61f0-4b75-9b7c-3fd51af35589	2.5.0	G431	Q235	Q235
1650e2f0-9a95-4422-9aa9-97397c1114d9	2.5.0	G431	Q235	Q235
29ee60cc-7c87-4cf0-acb9-37fca9c014db	2.5.0	G431	Q235	Q235
c4f0de97-f99a-4775-a3ff-eaa7a80051d7	2.5.0	G431	Q235	Q235
d7275905-c038-4ce6-bf58-66efaa6e399d	2.5.0	G431	—	Q235
4549f6d9-7342-47c4-a329-2af2ab244006	**UNVERIFIED**	**UNVERIFIED**	**UNVERIFIED**	**UNVERIFIED**
978dbf3d-82c3-4286-a233-f0029b606440	**UNVERIFIED**	**UNVERIFIED**	**UNVERIFIED**	**UNVERIFIED**
e12c6b39-1ac3-48d4-a77d-dfdbcfe82848	**UNVERIFIED**	**UNVERIFIED**	**UNVERIFIED**	**UNVERIFIED**

Table 15: Exact UUIDs and recorded model fields for runs addressed by the dated analysis artifact (part 3). UNVERIFIED rows are named by the analysis source but absent from the hash-pinned publication export.

B Anti-Gaming Design (Rationale Level)

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

B.1 The gaming problem

Let A denote the set of AI mediation systems and $s(a, r)$ the score of agent $a \in A$ on run r. Traditional leaderboards rank by maximum score, $\text{Rank}(a) = \max_{r \in R_a} s(a, r)$, which is manipulable: an agent improves its apparent rank through selective publication without improving quality. Running n evaluations and publishing $\arg\max_r s(a, r)$ dominates honest reporting whenever score variance exists — which is inevitable under stochastic evaluation. (Benchmarking differs structurally from voting, where Gibbard–Satterthwaite impossibility applies to strategic voters [28, 29]; here the *publisher* is the strategic actor.)

B.2 Median-of-N scoring

The platform's leaderboard ranking rule is median score over a minimum of N runs:

$$\text{Rank}(a) = \text{median}(\{s(a, r_1), \ldots, s(a, r_N)\}). \tag{8}$$

Many snapshot cells have one replicate, and the replicated Qwen 2×2 has two per arm; no result reported here reaches the minimum N=5 — extension rows in particular are provisional and unranked until they accumulate N=5 runs. The rule therefore does not bind this paper's results and instead governs how the public leaderboard absorbs repeated runs as they accumulate.

Theorem 1 (Median-threshold robustness, odd N)**.** *For ranking by median over an odd number N of runs (minimum N=5), the median strictly rises above the previous median if and only if at least $\lceil N/2 \rceil$ resulting scores are above the previous median.*

Proof. For odd N the median is the $\lceil N/2 \rceil$-th-ranked value when sorted (N=5: the 3rd). A strict increase above the old median requires that order statistic to exceed it, equivalent to at least $\lceil N/2 \rceil$ resulting scores above the old median. The raw count of improved runs is not sufficient: $[10, 20, 30, 40, 50] \to [10, 21, 30, 41, 51]$ improves 3 runs but leaves the median at 30, while $[10, 20, 30, 40, 50] \to [10, 35, 30, 40, 50]$ improves 1 run and raises the median to 35 because 3 resulting scores exceed the old median. Unlike maximum-based ranking, where one outlier suffices, median-based ranking rewards moving the distribution past the threshold. □

For even N (reachable as runs accumulate past the minimum) the median is the midpoint of two order statistics and the equivalence fails in one direction: at least $\lceil N/2 \rceil$ scores above the old median no longer guarantees a strict rise (with previous median 30, the set $[10, 20, 29, 31, 32, 33]$ has three values above 30 and median exactly 30). The manipulation-resistance direction — a strict rise still requires moving at least half the distribution past the old median — is unaffected; the biconditional is claimed for odd N only.

Corollary. Median ranking limits the influence of a single outlier in either direction and is statistically robust in that narrow sense [30]. It prevents cherry-picking only if the registry admits all eligible completed runs under a fixed inclusion rule. A strategic publisher who can run many trials and select which N enter the registry can still manipulate a median. The platform records an internal append-only audit trail, but the withheld enforcement records prevent external verification that no eligible run was omitted; no stronger anti-selection claim is made here.

B.3 Evaluated fixtures and sequestered reserve

The authored fixture inventory contains 62 evaluated/development fixtures and 38 sequestered reserve fixtures. The reserve has not been run. The evaluated fixtures also informed benchmark development, and scenario-level records are not public. The current results therefore characterize performance on the evaluated fixtures; they do not establish contamination-free generalization or performance on the reserve. If the reserve is activated, it will be run under a locked protocol and reported as a new evaluation epoch.

B.4 Canary strings and audit trail

Each scenario embeds a unique per-scenario canary string; a canary appearing in model output provides probabilistic evidence of scenario-text leakage. Construction and detection details are not disclosed. An append-only audit trail supports an internal provenance audit of benchmark runs. Because the scenario-level records and enforcement mechanics are withheld, this control does not enable independent third-party verification of the full result chain.

B.5 Design caveats

The minimum-run requirement delays new entrants' visibility; the unrun reserve is finite; and canary detection is probabilistic, not deterministic.

C Secondary Judge Probes

Two secondary probes are recorded here because the body text declares both non-decisive: the first sits outside the pre-stated panel rule, and the second is superseded by the fixed-transcript crossed re-judge.

Post-hoc fifth crossed judge. A post-hoc fifth crossed judge, added after the verdict to probe the GLM-5.2 anomaly, used GLM-4.7 served on Cerebras rather than GLM-5.2 through the HF router. It completed 557 of 558 evaluations and rank-correlates with the operational, frontier, gpt-oss, and Kimi judges at Spearman 0.68–0.79, while correlating at 0.17 with GLM-5.2 itself; on the adversarial ordering it votes with the majority (+4.0 [−4.0, 12.5]). Because both model version and serving stack changed, this comparison is only consistent with model-version and/or serving-stack sensitivity; it does not isolate a cause or clear the model family. The 2–2 verdict of Section 3.4 was prospectively specified over the original panel and stands; this paragraph is a post-hoc sensitivity probe, not a re-vote.

Judge-swap generations (context only). Independent baseline *generations* under a different operational judge differ by 2 points under the adversarial participant (18 vs. 16) and 6 points under the cooperative participant (36 vs. 42). These judge-swap rows are separate conversations, not re-scores — judge influence is confounded with generation stochasticity — and are superseded by the crossed re-judge of Section 3.4, which isolates scoring differences on fixed transcripts.

D 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 / time-to-agreement.** T_s = first participant turn with agreement or judge 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-agreement 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).** In the *full bilateral brief* condition, mediated arms receive

both parties' researcher-authored private scenario facts verbatim, and judges see both parties' backstories and stated goals (Section 2.4). In the *public-context-only* condition, the mediator receives the public scenario description, shared facts, and conversation, but not either party's private facts. This manipulation is a controlled analogue of differences in intake or caucus preparation; it does not observe or simulate a human preparation interview. The frozen prospective design is a model-matched 2×2 (policy $\times$ information context) on the adversarial participant: Reflective Listening and Skilled policies both bound to the *same* Qwen backbone, each run under both information conditions. Model matching is required because the frozen snapshot's active control ran on an unpinned harness-default backbone (Section 4). Estimands, frozen with the analysis code before data (Section 2.8): the public-context skill increment E_1, the full-brief skill increment E_2, the interaction $E_3 = E_1 - E_2$, and per-policy information-context contrasts; four-arm scenario-co-resampled bootstrap over fixed observed per-scenario replicate means. Reading rules: the within-harness skill contrast remains positive under public-context-only information iff E_1's interval excludes zero from below; the increment survives *without material attenuation* only if E_3's interval additionally clears a prospectively frozen non-inferiority margin of 4.0 composite points — an interaction interval straddling zero is not evidence of equivalence. One disclosed asymmetry is part of the estimand: mediator prompts carry a sliding window of the last 20 messages, so a full-brief mediator is re-handed the briefing every call while a public-context-only mediator's elicited discoveries can scroll out. *Delivered post-freeze (2026-07-21):* the Qwen 2×2 ran with two replicates per arm and resolves its prospectively frozen E_1 and E_3 decision rules affirmatively (Section 3.6). Conditional breadth runs (Kimi and gpt-oss backbones, added 2026-07-17, two days after the estimand freeze, as an outcome-independent amendment) show that Kimi crossed both frozen numerical thresholds while gpt-oss crossed neither; only Qwen is a replicated core. A post-hoc assay-sensitivity safeguard added 2026-07-26, after those outcomes were known, withholds Kimi's E_3 interpretation because its E_2 interval includes zero. That safeguard did not change an estimand, margin, threshold, or stored frozen-rule result. A separately frozen evaluator-context re-score was registered on 2026-07-22 before any context-withheld score existed, using the pre-existing context-visible scores as its observed reference. It completed on 2026-07-31 at exact scenario-set coverage for all 9 runs. The canonical aggregate reports the frozen context-withheld-minus-context-visible per-run and per-contrast estimands only at aggregate level under the benchmark's anti-gaming disclosure policy; it did not change any primary estimand, margin, or decision rule.

10. **Exploratory Gemma participant-transport extension.** The initial Skilled-policy extension was frozen on 2026-07-24 before either mediated Gemma run. On 2026-07-25 at 04:19 UTC, while both Skilled cells were still generating and the admin record showed 0/62 scenarios evaluated, zero failures, and no mediated score for either run, the design was expanded outcome-independently to the complete model-matched 2×2: Reflective Listening and Skilled policies, both bound to Qwen3-235B, under full bilateral briefing and public-context-only access. This is a post-launch, pre-outcome exploratory amendment, not part of the original Stheno factorial freeze or the initial Gemma freeze. All four mediated cells are newly generated under suite 2.5.0; the completed Gemma-4-31B no-mediator run supplies the matching baseline but is not a factorial arm. The existing E_1–E_4 definitions (Section 2.8), bootstrap, seed, and 4.0-point interaction margin are applied unchanged. With one run per arm, the readout is conditional single-replicate participant breadth, not a run-sampling estimate. Relative to the Stheno matrix, the extension holds mediator backbone, policies, information manipulation, judge, scenario partition, suite, and launch epoch fixed within Gemma, partially reducing

apparatus confounding. It does not isolate disposition: participant backbone, training and alignment, disposition, serving provider, and unseeded generation still change together, and the prospectively specified same-model persona swap remains open. The baseline and four mediated runs were kept private through completion, cleaning, and provenance checks, then released together regardless of result direction.

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 [16]; a shared global conversation budget and symmetric stopping ablation, informed by ProMediate [17]; and participant-composition, history, emotional-reactivity, and cultural stress axes with expert evaluator calibration, informed by SoCRATES [18]. 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. (4) 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.