MediationBench: Auditing AI Mediation in Synthetic Conflict

Jonathan Hendler*
HAI.AI
https://mediationbench.com/

Working draft — analysis snapshot 2026-07-26. Numbers may change between revisions.

Abstract

We present MediationBench, a dyadic benchmark that scores *how* simulated disputes evolve — not merely whether they end — using a composite 0–100 HAI Score (five expert-weighted rubric components) over paired mediated and no-mediator runs of 62 evaluated conflict scenarios in a fixed development partition. The frozen gate matrix ran under suite 2.0.0 (methodology seeded_25_v1); the information-study cells ran under suite 2.5.0, a mixed provenance stamped wherever those cells are reported. Changing the participant simulator changes the instrument's observed dynamic range: skilled-mediator point scores occupy the narrow 54–55 band under a cooperative assistant-aligned participant but span 44–48 under a higher-resistance open-weight participant. Because these configurations jointly change model capability, disposition, and alignment, this is a benchmark-validity result, not a causal claim about a single participant property. A minimal Reflective Listening active control also recovers much of the measured gain (+22 of +30 in the higher-resistance column), partly because a mediator-only rubric component mechanically rewards any mediated arm.

Our strongest prospective contrast is model-matched and ran on the adversarial (Stheno-v3.2) participant column: with Qwen and public-context-only information, the Skilled policy exceeds the model-matched Reflective Listening active control by $E_1 = +13.17$ [9.97, 18.61] composite points (90% scenario-reweighting interval, conditional on the observed runs and operational judge). The replicated Qwen core passes both frozen decision rules — E_1 positivity and interaction non-inferiority (margin -4.0 points). Kimi, run mostly once per cell, also clears both rules, but its full-brief increment interval includes zero, so its non-inferiority reading lacks assay sensitivity; gpt-oss clears neither. Separately, the Skilled-policy public-minus-full contrast $E_{4,\text{skilled}}$ clusters near zero in point estimate across the three tested backbones; every interval includes zero. We did not prespecify an equivalence margin for that estimand, and the full-brief Skilled arms were generated earlier under an older suite version, so we read the pattern as hypothesis-generating, not evidence of equivalent information conditions. The full-bilateral-brief versus public-context-only manipulation is a controlled analogue of preparation and intake, not an observed human interview or a test of human mediation efficacy.

We *define* paired effectiveness deltas with a 90% scenario-reweighting interval (scenario-cluster bootstrap), a per-turn cooperation curve with explicit survivorship conditioning, and a judge-reliability protocol (intra-judge test–retest and crossed second judges); the headline estimate is $\Delta_{\text{med}} = +36.9$ HAI Score points (90% CI [31.9, 41.4]; n=62 scenario pairs) for the strongest gate cell (Skilled@Qwen3-235B under Stheno-v3.2), and +35.5 [30.5, 38.0] for the same cell re-scored by the crossed judge. The crossed frontier judge reproduces the full mediator ordering on identical transcripts. A four-judge cross-family panel splits 2–2 on the ordering within the

*Corresponding author: jonathan@hai.io.

skilled pair — judge-indeterminate by a rule stated before the two added judges produced scores — while the arm-versus-control structure survives the panel's high-agreement judges (scenario-level $r = 0.84$, $\rho = 0.80$, $n = 558$, with the crossed frontier judge and $r \approx 0.67$ with the open-weight panel judge). Intra-judge test–retest yields Krippendorff's α of 0.48–0.91 by component. Four targeted end-to-end repeats show a maximum run-level movement of 3.7 composite points while the within-pair point ordering flips; they do not bound run-to-run variation. These results are descriptive for a fixed harness — mostly one to two runs per cell, two confounded participant configurations, and a single operational judge — and the prospectively specified analysis plan states what stronger claims will require. Frozen gate aggregates are versioned in a public export; the export does not contain the scenario-level records needed to reproduce the resampling intervals. A single-run Gemma participant replication cleared the baseline in every mediated arm yet failed both frozen rules: the public-context skill finding did not transport.

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 mediation-quality score (the HAI Score, five expert-weighted 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. This contrast shows that participant configuration materially changes the benchmark's ability to distinguish mediator packages; it does not establish which property of the participant model causes that change. The pre-declared paired within-column contrast supports the claim directly: the within-skilled-pair contrast is +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 A. 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 questions that end-state scoring conflates: whether adding any mediator changes the conversation (paired baseline contrasts), whether a skilled policy adds value beyond minimal reflective listening (E_1, E_2), whether supplying both parties' private brief changes what the mediator can accomplish (E_3, E_4), and whether a conclusion survives a different run or evaluator (the replication and judge-reliability 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, a validated mediator-presence detector with a measured resolution limit: every frozen gate-cell interval excludes zero and the headline contrast survives the crossed frontier judge (a judge that shares a model family with the control's harness-default backbone, Section 2.4), while skilled arms collapse into one tie band per participant column — a citable bound on what composite conversation-level scoring can currently distinguish. Second, a reusable evaluation asset — scenario suite, transcripts, judge rubric, and versioned score data — that can support future evaluation and simulation studies.

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 make different failures observable, 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).

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 cancels the shared baseline. 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:

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 A); 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 A).

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 (Section 3.4) — 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 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 composite-score improvement 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 reported agreement 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.

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 controlled analogue of preparation through intake or caucus. 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 (Section 5) crosses policy and information within each backbone.

Judges (the operational judge and every offline panel judge, whose input is byte-identical by construction) receive 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, a private fact can enter the conversation only if a participant voices it (Section 5).

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, creating greater dynamic range for 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.

2.6 The HAI Score: what we measure

The HAI Score (0–100) is a composite metric designed to evaluate cooperative transformation quality in mediation; Table 2 lists its components and weights, which Eq. (1) combines.

Component	Weight	What it measures
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 expert-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.

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). Concretely: a disputant whose scenario fixture assigns the hidden constraint *"if the partnership dissolves, I lose my immigration status"* is *hidden* while that fear shapes refusals nobody can explain, *hinted* when they allude to unspecified personal reasons for urgency, and *revealed* when the constraint is on the table and the parties can problem-solve around it. Breakthrough moments in conflict resolution frequently occur when hidden interests behind stated positions become explicit [22, 23], so a mediator that helps place a constraint safely on the table can advance collaborative problem-solving. The information condition (Section 2.4) scopes that interpretation: under the full bilateral brief, revelation measures whether known hidden state — supplied verbatim in the mediator's packet — is 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 test–retest distribution shows the judge detects revelation while the deterministic answer-key cap removes it (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 (Section 5).

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

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

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 5). A scenario marked *failed* has no score: $S(s)$ is undefined and s is excluded from all pairing below.

Matched baseline. The baseline cohort of cell c contains exactly those baseline (no-mediator) runs that are public, completed with at least one scenario evaluated, contain *no* failed scenario, and agree with the mediated run on all four matching keys: suite version, participant model, judge model, and the exact scenario set $\mathcal{S}$. Pooled cross-model baseline statistics are never used; the comparison

is always against this matched cohort. When either arm contains more than one qualifying run, the per-scenario score of that arm is the continuous median across its runs. (Replicate runs are therefore analyzed one at a time when reported as separate values; pooling replicates into one cohort would silently average them.)

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,

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

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), replicates are resampled with replacement within each sampled (scenario, arm) and re-averaged, and every draw recomputes the same functional as the point estimate.

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 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 (Section 5). One reporting precondition guards assay sensitivity: an E_3 non-inferiority verdict is reported only where that backbone's E_2 interval excludes zero, since otherwise the margin permits attenuation larger than the reference effect it protects.

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

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

with q_p the bare order statistic at the p-th index of the sorted draws (no interpolation). Because pairs are co-resampled, the two arm medians share the resampled scenario set in every draw, so the interval propagates the positive within-scenario correlation between arms.

The bootstrap RNG is Python's Mersenne Twister (`random.Random`) with fixed seed 20260706, so the computation is exactly reproducible; this seed governs resampling only and is unrelated to LLM token sampling, which is unseeded. The same seeded stream is re-created for every reported interval, so Monte-Carlo resampling error is common across intervals rather than 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 shared Monte-Carlo resolution. Where a frozen estimand aggregates replicates inside each bootstrap draw (Section 2.8), replicate resampling is nested within scenario reweighting; the resulting interval still conditions on the finite set of observed generation runs.

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 (Section 5).

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 confirmatory primary endpoint; 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 on ordinal buckets is conservative; 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 agreement 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.

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

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

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. 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 robustness 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.

3.1 Effectiveness deltas 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 confirmatory statement about the frozen family; post-freeze extensions are labeled exploratory where they appear. The benchmark's central structural observation is that **the tested participant configuration changes the instrument's observed dynamic range.** 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, so these data do not identify which property causes the difference.

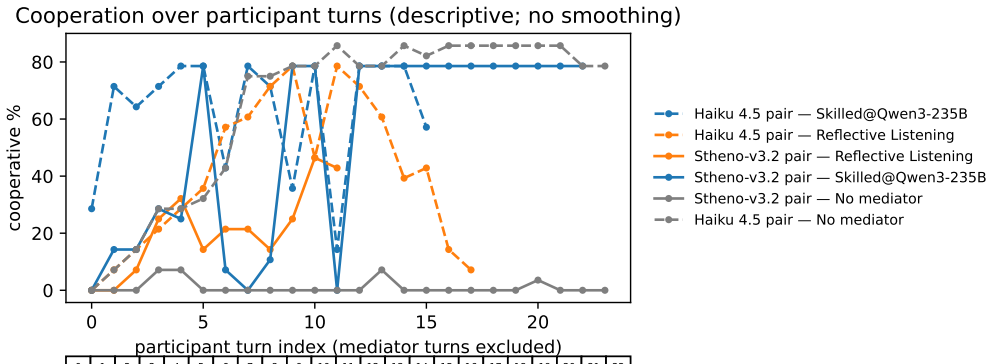
Mediation raises measured cooperation in every published mediated cell, with larger absolute deltas under the adversarial participant (+30 and +26 versus +18 and +18). The same holds at scenario grain once the harmed scenarios are counted, per the flagged-not-suppressed rule of Section 2.3: between 0–8 of 62 scenarios per gate cell score *below* their matched baseline — Skilled@Qwen3-235B under Stheno-v3.2 (L3-8B) worsens 0 scenarios, while the Reflective Listening control under Claude Haiku 4.5 worsens 8. These per-cell counts should be read against the generation-noise floor measured by the baseline re-runs: re-running the identical no-mediator configuration moved 56 (adversarial) and 57 (cooperative) of 62 scenarios in one direction or the other (Section 3.5). That noise floor is the only available bound on how much of the harm statistic is arm effect rather than generation noise. 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 conversational improvement. 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 observed conversational change under interleaved reflection, visible directly in the Cooperative Dimensions component (Section 3.3). The quantity the benchmark *ranks* 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). 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.

Three 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 (Section 5). This active-control decomposition is therefore a decomposition of the mean gain, not a significance claim about any mediator pair.

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



	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.

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.

Mediation also changes how long conflicts run. Baseline conversations 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 (Section 5, 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 4 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 expert-asserted,

and it motivates a weight-sensitivity analysis — a Moderator Quality-excluded, renormalized delta co-reported as a labelled sensitivity, never as a replacement for the frozen primary estimand — which we register as future work in Section 5; it was not in the frozen plan. 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 4). Fourth, Hidden-State Revelations, the construct's nominal 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), so the judge detects revelation and the deterministic answer-key overlap cap removes it; component-level conclusions there are unsafe.

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

3.4 Judge reliability

Crossed re-judge (identical transcripts). Two 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 (Section 5) 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 5). 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 agreement is scenario-level $r = 0.84$, $\rho = 0.80$, $n = 558$, with the crossed frontier judge and $r \approx 0.67$ with the open-weight panel judge. Median deltas under all three judges appear in Table 5; 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 confirmatory statement — 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 self-preference account of the ordering finds no support on either side. Second, GLM-5.2 is a measured outlier. Its scenario-level Spearman agreement 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 effects, and sign-reverses the ranked skilled-over-control quantity itself (−9.0 [−14.0, 4.0], interval spanning zero). We discount that reversal on the measured-agreement grounds above, not on its direction. Its missing evaluations also concentrate in the longest, arm-correlated transcripts, and the panel rule counts unweighted signs. Under any reliability floor that would exclude a judge agreeing 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 effects are what survive the panel's high-agreement judges.

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 agrees 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 above was prospectively specified over the original panel and stands; this paragraph is post-hoc robustness, not a re-vote.

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.

Participant	Arm	Δ_{med} [90% CI] by judge		
		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 5: 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).

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 reflects restricted range in this sample rather than 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 published coefficients are those retained by the calibration snapshot, and 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, Section 5, 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 — measured, moderate noise. 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 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 effect is confounded with generation stochasticity — and are superseded by the crossed re-judge above, which isolates scoring differences on fixed transcripts.

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 confirmatory 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.

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, no-CI rows with only a few points of separation; the observation is a hypothesis-generating one that the prospectively specified replicated design (Section 5) can test, namely whether mediation quality is a capability 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 → 37.9, adversarial baseline 18 → 15.5, Skilled@Kimi-K2.5 44 → 46.2, Skilled@Qwen3-235B 48 → 44.3. At scenario level the repeated mediation deltas differ by at most 1.2 points from the frozen values (+35.8 [30.2, 37.6] vs. +36.9 [31.9, 41.4] for Skilled@Qwen3-235B; +35.1 [29.4, 38.7] for Skilled@Kimi-K2.5). The frozen-minus-repeat baseline intervals include zero (+0.3 [−1.0, 3.5] cooperative, −0.4 [−5.5, 2.2] adversarial), 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 baseline repeats also expose the generation-noise floor: the identical no-mediator configuration moved 57 (cooperative) and 56 (adversarial) of 62 scenario scores in one direction or the other, the reference against which the per-cell harm counts of Section 3.1 must be read. 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 mediation-effect 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 (Section 5) 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 6. 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 $E_1 = +13.17$ [9.97, 18.61] 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 effectiveness in human disputes. Under full bilateral briefing, the Skilled-policy increment is $E_2 = +7.27$ [1.46, 12.86]. The interaction is $E_3 = +5.9$ [0.56, 14.62] 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 [−15.18, −2.55] and the Skilled policy is +0.01 [−5.40, 2.69]; 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 [9.97, 18.61]	+7.27 [1.46, 12.86]	+5.9 [0.56, 14.62]	−5.89 [−15.18, −2.55]	+0.01 [−5.40, 2.69]
Kimi	2/1/1/1	+9.66 [2.32, 16.46]	+1.68 [−1.06, 6.79]	+7.98 [−1.09, 14.50]	−6.45 [−12.10, 0.58]	+1.53 [−2.17, 4.24]
gpt-oss	1/1/1/1	−1.02 [−4.51, 5.01]	+4.74 [2.15, 15.53]	−5.75 [−16.63, 0.32]	+4.12 [0.33, 15.10]	−1.64 [−4.57, 3.38]

Table 6: 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 [−5.40, 2.69], +1.53 [−2.17, 4.24], and −1.64 [−4.57, 3.38]). Each of these three $E_{4,\text{skilled}}$ intervals includes zero. 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 [−2.17, 4.24]) exceeds it only at the upper end, while Qwen's (+0.01 [−5.40, 2.69]) and gpt-oss's (−1.64 [−4.57, 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 [−15.18, −2.55] on Qwen but +4.12 [0.33, 15.10] 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.

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 supports the E_1 and E_3 decision rules, but its E_2 interval (+1.68 [−1.06, 6.79]) includes zero, so by the assay-sensitivity precondition of Section 2.8 its non-inferiority reading is not reported as a verdict: the margin there permits attenuation larger than the reference effect it protects. gpt-oss fails both rules: the Skilled policy does not beat its public-context-only Reflective Listening control and the interaction does not clear the margin, while that control improves without the private-facts packet. Only the Qwen matrix has two runs per cell, so Kimi is conditional support 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). Mediation transported: every mediated arm cleared the scenario-matched no-mediator baseline (Table 7), and the full-brief skill increment was clearly positive ($E_2 = +11.39$ [5.77, 16.79]). The public-context skill increment was not: $E_1 = +5.70$ [−3.38, 12.89] fails the frozen positivity rule, and the interaction $E_3 = -5.68$ [−15.55, 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 effect was flat ($E_{4,\text{reflection}} = +0.73$ [−3.58, 6.23]) and the skilled policy's uncertain ($E_{4,\text{skilled}} = -4.95$ [−13.69, 1.96]). 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 6.

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 7: Participant-transport replication (registered extension): Gemma-4-31B participant pair under the Qwen mediator backbone, one run per arm. Δ_{med} is the median of per-scenario differences from the scenario-matched no-mediator baseline, with 90% scenario-reweighting intervals. Every mediated arm clears the baseline; the frozen E_1 and E_3 rules nevertheless fail on this column (see text).

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 Section 5.

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 analyzed configurations show that participant configuration changes what this harness measures; they support no pooled population-level claim. One registered extension covers the participant axis: the exploratory Gemma participant-configuration replication of the information factorial (Section 5, 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 do not identify which participant property causes a dynamic-range difference. The same-model persona-swap deconfound remains prospective.

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 information factorial is not wholly contemporaneous.** For each backbone, the public-context-only policy contrast (E_1) and the reflective-policy information contrast use contemporaneously generated suite-2.5.0 cells. The full-brief policy contrast (E_2), the interaction (E_3), and the Skilled-policy information contrast combine those cells with protocol-matched frozen suite-2.0.0 full-brief Skilled-policy arms. Run date and suite stamp therefore differ; historical drift or other version-linked differences can affect those three estimates (argued 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.

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 effects.

10. **Component weights are expert-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; per-component results and any mechanical decomposition remain exploratory.

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, or judge scores.

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 effectiveness of mediator preparation, interviewing, or mediation. 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 Frozen Prospective Analysis Plan and Future Work

This section 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 effects; four-arm scenario-co-resampled bootstrap with replicate resampling. Reading rules: skill remains beneficial 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) give support on Kimi but fail to generalize on gpt-oss; only Qwen is a replicated core. A separately frozen evaluator-context re-score has been executed. Context-withheld judge rows appear in the generic crossed-judge output, but the prospectively frozen matched context-withheld-minus-context-visible summary has not yet been emitted as its own canonical aggregate. This version therefore does not promote it to a paper estimand; once emitted, it will be reported only at aggregate level under the benchmark's anti-gaming disclosure policy.

10. **Exploratory Gemma participant-configuration replication of the policy-by-information factorial.** 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 remain private until all five are complete, clean, and provenance-checked, then are made public 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).** 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.

6 Conclusion

We presented MediationBench, a benchmark that measures whether an AI mediator improves a simulated dyadic dispute, 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 configuration changes the instrument's observed dynamic range: against the cooperative configuration, skilled-mediator point scores occupy a narrow band, while the tested higher-resistance configuration produces a wider spread. Capability, disposition, and alignment are confounded between those participant configurations, 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 improves the measured outcome, 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 preserves arm-versus-control structure but not the fine ordering within the skilled tie band, which a prospectively specified panel rule records as judge-indeterminate. 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 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 $E_1 = +13.17$ [9.97, 18.61] composite points (90% scenario-reweighting interval conditional on the observed runs and operational judge). Kimi gives conditional, mostly single-run support while gpt-oss clears neither frozen rule, so the decision-rule outcome does not generalize across the three tested backbones. 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, 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 practical implication for evaluation is concrete: score 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 participant configurations, 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.

Two next steps are planned. First, human expert review of transcripts and judgments, giving the composite its first criterion-validity check (the frozen plan's reliability-extensions item — human-vs-judge calibration — is the reporting slot). Second, broadening the judge panel and the participant configurations to models from more foundation-model companies, completing MediationBench's coverage on the two axes this paper shows matter most.

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. No institutional review was required because no human subjects, human-derived transcripts, or personal data were involved. 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, and no harmful-content generation beyond scenario-bounded dispute behavior is elicited or published. We note the dual use plainly: measurement machinery that identifies which conversational moves de-escalate can be inverted to select the moves that inflame, and this inversion is not hypothetical — a major platform experimentally manipulated users' emotional states through feed curation [24], and engagement-optimized ranking systematically amplifies out-group animosity [25]. Participants capable of sustained hostility remain necessary regardless: a benchmark cannot measure a mediator's ability to reduce toxicity 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 their score-level 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 measures score-level harm only.

Reproducibility and Data Availability

Published results are aggregate and versioned. The public export (https://mediationbench.com/results/) carries, for every run: the exact participant, mediator, and judge model identifiers, suite and methodology version stamps, run-level scores and deltas, and per-component category scores; per-role sampling temperatures are recorded once at export level (shared across runs), and per-turn cooperation-curve aggregates with per-point n are present for most but not all archived runs. Sampling is unseeded and disclosed as such.

The export supports inspection and recomputation of the run-level aggregates it contains, 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 committed alongside the paper (`tools/compute_claims.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_2026-07-26.json`); only the export snapshots themselves are hash-identified. 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 suite-2.5.0 information-study generation run — including the pooled full-brief repeat — 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. 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 the paper's numbers remain checkable after the live export moves on; 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 adversarial participant's model identity is pinned; its unseeded behavior is stochastic rather than exactly reproducible. Assistant-aligned participants are identified by exact API model version. The adversarial participant is a Llama-3-8B derivative and is used under the Meta Llama 3 Community License ("Built with Meta Llama 3").

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: archive the exact export snapshot (Zenodo DOI recommended) and add the exact-model-id pinning table as an appendix (API model ids and dated snapshots, HF provider/revision for open-weight models, scenario corpus hash, judge prompt version).

Competing Interests

The benchmark, harness, and mediator implementations evaluated here are developed by HAI.AI, which operates the hosted mediation product the benchmark scores among others. Active mitigations in the published data: paired baselines and a minimal Reflective Listening active control that any Skilled policy must beat, 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 and recompute run-level aggregates. Vendor diversity is not independent third-party evaluation. Both prospectively frozen internal judge-sensitivity checks have been executed. The crossed-judge analysis is reported in full (Section 3.4); the evaluator-context ablation is not promoted to a paper result because its matched context-withheld-minus-context-visible summary has not yet been emitted as a dedicated canonical aggregate (Section 5).

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, a prospectively specified 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. *arXiv preprint arXiv:2410.07053*, 2024.

[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. *arXiv preprint arXiv:2509.06586*, 2025.

[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. *arXiv preprint arXiv:2410.02736*, 2024.

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

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

[24] 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.

[25] 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.

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

[27] 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.

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

A 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.

A.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 [26, 27]; here the *publisher* is the strategic actor.)

A.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. Random variance has limited effect on median-based rankings; the median discards outliers in both directions, making the ranking robust in the statistical sense [28]. Median-of-N applies to baseline runs as well, so neither arm of the delta rewards cherry-picking.

A.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.

A.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.

A.5 Design caveats

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