

SocialCritic: Grounding Multi-Agent Dialogue via Advantage-Based Endorsement Suppression

Shuang Cao* and Rui Li*,†

Hill Research

619 Alexander Road, Princeton, NJ 08540, USA

sharon.cao@hillresearch.ai rui.li@hillresearch.ai

*Equal contribution. †Corresponding author

Abstract

In multi-agent dialogue, a hallucination crosses an operational control boundary when another agent accepts it and downstream decisions inherit the error. We formalize this cross-speaker transition as *unsupported endorsement* and introduce SOCIALCRITIC, a learned receding-horizon controller that ranks and repairs candidate action sequences with separately trained utility, norm, and epistemic advantage critics. Across 692 test episodes spanning LIGHT, DealOrNoDeal, SOTOPIA, and template-family-held-out Persuasion-Conflict over 10 generation seeds, SOCIALCRITIC exceeds SAFE-Dialogue by 6.9 points in goal completion and 7.9 points in actionable accuracy while reducing endorsement by 44.9% and norm violations by 42.7%. The public-benchmark aggregate preserves the actionable-accuracy gain (74.9% vs. 66.8%), and trace analysis shows that endorsement falls 77.5% versus 63.5% for initial injection relative to LLM-only. Leave-one-benchmark-out critics retain an actionable-accuracy advantage on every target (minimum 5.2 points), while an $H=1$ controller without partner rollouts preserves 71% of the full controller’s advantage with 18% lower latency. At the default operating point, batched local scoring requires 1.7 external round trips and 748 ms per turn, enabling propagation-aware simulation, safety evaluation, negotiation, tutoring, moderation, and red-team workflows.

1 Introduction

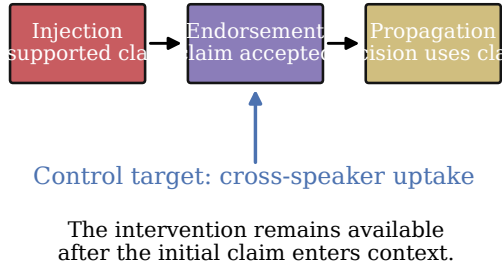
Multi-agent LLM systems are increasingly used to study negotiation, social interaction, and collaborative reasoning (Park et al., 2023; Piao et al., 2026). Their distinctive reliability risk is social propagation: one agent injects an unsupported assertion, another accepts it, and the dialogue turns that acceptance into a decision. For example, Agent A

states “the venue confirmed March 15” without evidence; Agent B replies “then I will finalize catering,” converting a hallucination into a commitment. The second turn marks an *endorsement boundary*: a distinct intervention point that remains controllable after the first claim has entered context. We call the boundary-crossing event *unsupported endorsement*, the cross-speaker uptake of a factual or social claim that the available state, evidence, and dialogue history do not support.

This construct connects hallucination to a trace-level consequence. Injection measures unsupported generation; endorsement measures whether another speaker accepts or builds on that content; propagation measures subsequent repetition or action. Across our four evaluated domains, SOCIALCRITIC reduces the claim-level endorsement transition by 77.5%, compared with 63.5% for injection, and removing its epistemic critic causes the largest measured endorsement increase. Controlling uptake therefore complements generation-time factuality and directly serves researchers and developers running multi-agent simulation, negotiation, tutoring or moderation sandboxes, and red-team evaluations where socially reinforced errors matter.

Figure 1 presents SOCIALCRITIC, a learned multi-objective, receding-horizon dialogue controller. At every turn, an LLM planner proposes K action sequences. Three separately trained critics estimate discounted sequence-level utility, norm, and epistemic advantages from short simulator rollouts; thresholded span feedback then directs up to two repairs before the controller executes the first action of the best rescored sequence. The decomposition makes factual support and social compliance inspectable controls rather than a single holistic judge score. It also changes the computational profile: candidate generation and local critic evaluation are batched, yielding 1.7 ± 0.4 external round

(a) Injection → endorsement → action



(b) Multi-objective receding-horizon control

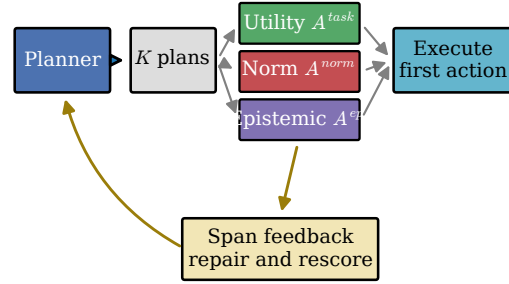


Figure 1: Unsupported endorsement and SOCIALCRITIC. Left: an unsupported injection becomes consequential when another agent endorses it and acts on it. Right: the controller generates K plan sequences, scores their simulated prefixes with utility, norm, and epistemic critics, repairs the highest-ranked risky sequence, executes its first action, and replans.

trips and 748 ± 62 ms per turn, versus 3.0 round trips and 1642 ± 214 ms for SAFE-Dialogue.

SOCIALCRITIC turns advantage-grounded planning into control at the endorsement boundary by jointly learning task, norm, and epistemic sequence signals with span-level repair. Relative to Best-of- N ’s external holistic judge, its decomposed critics are local; relative to evidence-constrained prompting and Self-Refine (Madaan et al., 2023), its repair signal is learned from downstream trace behavior; relative to post-hoc filtering, selection jointly advances the goal and reduces propagation before action execution. Persuasion-Conflict isolates the mechanism by crossing conflicting evidence and social incentives under template-family holdout, category balance, evidence-sufficiency checks, and explicit injection–endorsement–propagation traces.

Contributions. (1) The endorsement boundary as a measurable trace-level control target, with a cascade protocol that separates unsupported injection, cross-speaker uptake, and downstream propagation. (2) A receding-horizon controller that combines separately trained utility, norm, and epistemic sequence advantages for candidate selection, targeted span repair, and fully rescored execution. (3) Persuasion-Conflict, a template-family-held-out evaluation design for evidence-grounded dialogue under conflicting information and social incentives. (4) Matched-budget, component, leave-one-benchmark-out, no-partner-rollout, operating-frontier, robustness, and planner-backbone evidence. Under canonical metric-specific micro aggregation, SOCIALCRITIC improves goal completion and actionable accuracy over SAFE-Dialogue

by 6.9 and 7.9 points while reducing endorsement and norm violations by 44.9% and 42.7%; the public-only ranking, cross-domain critic transfer, and $H=1$ results preserve the central advantage.

2 Related Work

LLM agents and multi-agent dialogue. LLMs have been repurposed as cognitive simulators with memory and decision modules (Park et al., 2023), and evaluated on dialogue tasks including fantasy role-play (Urbanek et al., 2019), negotiation (Lewis et al., 2017), and social interaction (Zhou et al., 2024). Large-scale frameworks expose emergent social behavior (Piao et al., 2026); multi-turn interaction also creates a propagation channel through which one agent’s unsupported content can alter another agent’s beliefs and actions. Multi-agent debate can improve factuality (Du et al., 2024), but treats interaction as an elicitation mechanism rather than measuring cross-speaker uptake as the controlled failure event.

Sycophancy, over-agreement, and unsupported endorsement. Sycophancy research studies models adapting answers to a user’s stated belief or preference, including preference-model mechanisms and synthetic-data interventions (Sharma et al., 2024; Wei et al., 2023). Unsupported endorsement targets a distinct transition: one agent accepts or propagates another agent’s unsupported factual or social claim over a multi-turn trace. This distinction changes both intervention and evaluation. The endorsement boundary remains actionable after an initial injection, and its evaluation follows cross-speaker transition events—uptake and downstream

repetition or action—across the trace.

Grounding, planning, and refinement. ReAd (Zhang et al., 2025) uses sequential task advantage to refine embodied plans. SOCIALCRITIC redirects this planning primitive to propagation control: three separately learned sequence scores represent task progress, norm compliance, and epistemic support, and critic-localized spans become targeted repair feedback before an endorsed claim can drive action. Best-of- N instead delegates holistic ranking to an external LLM judge. Self-Refine (Madaan et al., 2023) uses unconstrained model self-feedback, while evidence-constrained and verification prompts apply fixed instructions. SOCIALCRITIC’s signals are trained from downstream trace behavior and applied under the same evidence scaffold and matched generation budget.

Factuality and dialogue safety. Hallucination research characterizes unsupported generation (Ji et al., 2023), and safety-oriented reward modeling shapes single-model behavior (OpenAI, 2023; Ouyang et al., 2022). Dialogue filters commonly reject outputs after generation; in contrast, multi-objective candidate control can improve task progress and groundedness together. Fact-checking benchmarks evaluate isolated claims against evidence (Su et al., 2024; Liu and Roth, 2025), not whether one agent’s unsupported statement is adopted and amplified by another during interaction. Persuasion-Conflict combines contradictory evidence, social incentives, template-family hold-out, and propagation traces, enabling direct comparison between controls that reduce initial generation and controls that prevent consequential uptake. Appendix B summarizes these method-level distinctions.

3 Method: SOCIALCRITIC

3.1 Problem Setting and Endorsement Boundary

We consider a (partially observable) social environment with latent state $s_t \in \mathcal{S}$. At each step, an agent produces an action $a_t \in \mathcal{A}$, where actions may include *utterances* (natural language) and *non-linguistic* actions (e.g., move, pick up, pass). The environment transitions as $s_{t+1} \sim P(\cdot \mid s_t, a_t)$. For claim c introduced by speaker q_t , let $u(c \mid h_t)$ indicate that the visible history and evidence do not support c , $m(x, c)$ be cross-speaker entailment, and $d(a, c)$ indicate a later repetition or action pred-

icated on c . Over the five-turn eligibility window $\mathcal{W}_t$ (or an earlier terminal action), the cascade events are

$$\mathcal{R}(c) = \{t' \in \mathcal{W}_t : q_{t'} \neq q_t, m(x_{t'}, c) > \delta_e\}, \quad (1)$$

$$\mathcal{D}(c) = \{t'' > t : d(a_{t''}, c) = 1\}, \quad (2)$$

$$E(c) = u(c \mid h_t) \mathbf{1}[\mathcal{R}(c) \neq \emptyset], \quad (3)$$

$$P(c) = E(c) \mathbf{1}[\mathcal{D}(c) \neq \emptyset], \quad (4)$$

where $\delta_e = 0.6$. $E(c)$ marks a claim crossing the endorsement boundary and $P(c)$ marks its downstream propagation; their eligible-claim and endorsed-claim denominators are specified in Section 4.3. The controller targets these transitions while preserving task progress and norm compliance through three reward channels:

$$r_t^{\text{task}} \in \mathbb{R} \quad (\text{task utility / goal progress}) \quad (5)$$

$$r_t^{\text{norm}} \in \{-1, 0\} \quad (\text{norm/ethics violation penalty}) \quad (6)$$

$$r_t^{\text{epi}} \in [-1, 0] \quad (\text{epistemic reliability penalty}) \quad (7)$$

We define a weighted return with discount $\gamma = 0.99$:

$$G_t = \sum_{k \geq 0} \gamma^k (r_{t+k}^{\text{task}} + \lambda_{\text{norm}} r_{t+k}^{\text{norm}} + \lambda_{\text{epi}} r_{t+k}^{\text{epi}}), \quad (8)$$

where λ_{norm} and λ_{epi} are hyperparameters controlling the trade-off between task success and constraint satisfaction.

3.2 Multi-Objective Sequence Critics

SOCIALCRITIC augments planner π_θ with task, norm, and epistemic critics indexed by $j \in \{\text{task}, \text{norm}, \text{epi}\}$. Each outputs a normalized step-level advantage proxy $\hat{A}^j(\hat{s}_{t+h}, a_{t+h}) \in [-1, 1]$. For candidate sequence $\mathbf{a}_{0:H-1}^{(k)} = (a_0^{(k)}, \dots, a_{H-1}^{(k)})$, let $A_k^j(s_t) \equiv A^j(\mathbf{a}^{(k)} \mid s_t)$. Each head is aggregated separately over the same rollout and discount:

$$A_k^j(s_t) = \frac{1}{Z_H} \sum_{h=0}^{H-1} \gamma^h \hat{A}^j(\hat{s}_{t+h}, a_h^{(k)}), \quad (9)$$

$$Z_H = \sum_{h=0}^{H-1} \gamma^h.$$

Normalization keeps all three sequence scores in $[-1, 1]$ for every horizon. The controller combines them as

$$A^{\text{total}}(\mathbf{a} \mid s) = A_{\phi}^{\text{task}}(\mathbf{a} \mid s) + \lambda_{\text{norm}} A_{\psi}^{\text{norm}}(\mathbf{a} \mid s) + \lambda_{\text{epi}} A_{\omega}^{\text{epi}}(\mathbf{a} \mid s). \quad (10)$$

Critic parameterization. Each critic is a three-layer MLP head over a frozen Llama-3-8B-Instruct encoder (Grattafiori et al., 2024). The task head regresses normalized TD(λ) returns with MSE. The norm head predicts violations under five categories (harm, deception, illicit guidance, privacy, and social contract), and the epistemic head predicts whether claims are unsupported or insufficiently calibrated. The classification heads use focal loss and output $1 - 2p(\text{violation})$ and $1 - 2p(\text{unsupported})$, respectively, aligning their semantics and scale with the task head.

Simulator interface. For $H=4$, each candidate prefix is advanced with the benchmark’s evaluation-time simulator and native scripted or learned partner-policy interface. The controller sees dialogue history, role metadata, retrieved evidence or context, and public simulator-returned state; it cannot access hidden partner goals, latent rewards, future labels, or test annotations. For $H=1$, Eq. (9) scores the candidate’s first action from the current visible state, makes no partner-policy call, and otherwise leaves planner, critics, prompts, thresholds, and candidate budget unchanged. This controlled variant isolates the value of lookahead from the value of learned critic guidance.

3.3 Receding-Horizon Inference

At decision time, the planner samples $K=8$ sequences with nucleus sampling ($p=0.9$, $T=0.7$). Algorithm 1 performs at most $R=2$ targeted refinements, executes only the first action of the best fully scored sequence, and replans from the resulting state.

Refinement feedback. CriticHighlights sentence-segments the selected sequence, scores every span with the norm and epistemic heads, and returns the lowest-scoring span from each head whose score crosses its threshold. Ties are resolved by earlier sequence position and then longer span. The planner receives only the quoted span, critic category, threshold margin, and a fixed-form edit directive (e.g., request evidence, calibrate certainty, or replace coercive language). When the task score

Algorithm 1 Receding-horizon SOCIALCRITIC inference

Require: Visible state s_t , planner π , critics $A_{\phi}^{\text{task}}, A_{\psi}^{\text{norm}}, A_{\omega}^{\text{epi}}$

Require: $K=8, H=4, R=2, \tau_{\text{norm}}=-0.1, \tau_{\text{epi}}=-0.1$

```

1: Sample  $\{\mathbf{a}_{0:H-1}^{(k)}\}_{k=1}^K \sim \pi(\cdot \mid s_t)$ 
2: for  $r = 0$  to  $R$  do
3:   Roll out and compute  $A^j(\mathbf{a}^{(k)} \mid s_t)$  for every  $j, k$ 
4:    $A^{(k)} \leftarrow A_k^{\text{task}}(s_t) + \lambda_{\text{norm}} A_k^{\text{norm}}(s_t) + \lambda_{\text{epi}} A_k^{\text{epi}}(s_t)$ 
5:    $k^* \leftarrow \arg \max_k A^{(k)}$ 
6:   if  $A_{\psi}^{\text{norm}}(\mathbf{a}^{(k^*)}) \geq \tau_{\text{norm}}$  and  $A_{\omega}^{\text{epi}}(\mathbf{a}^{(k^*)}) \geq \tau_{\text{epi}}$  then
7:     break
8:   end if
9:   if  $r < R$  then
10:     $f \leftarrow \text{CriticHighlights}(\mathbf{a}^{(k^*)}, s_t)$ 
11:    Replace  $\mathbf{a}_{0:H-1}^{(k^*)} \sim \pi(\cdot \mid s_t, f)$ 
12:   end if
13: end for
14: Return  $a_t^* = \mathbf{a}_0^{(k^*)}$ ; observe  $s_{t+1}$  and replan
```

is below the median candidate score, a one-line goal-progress shortfall is included; otherwise utility feedback is omitted. By rescoreing every current candidate, the next loop iteration ensures that the last permitted repair is evaluated before execution.

Call accounting. The eight initial sequences are generated in one batched request and count as eight generations but one external round trip. Each repair adds one sequential round trip, for at most ten generations and three round trips; threshold early exit yields 8.7 ± 0.4 generations and 1.7 ± 0.4 round trips per turn. All critic scoring is local and batched.

3.4 Critic Training

Training trajectories combine LLM-only, utility-guided, and randomized policies across the four designated benchmark training splits (44,600 episodes total). GPT-4o supplies five-shot weak supervision; 2,400 norm labels and 1,800 epistemic claim labels from automated reviewer profiles calibrate the classification heads, with agreement $\kappa=0.83$ and $\kappa=0.79$. The heads use AdamW (learning rate 2×10^{-5} , weight decay 0.01), batch size 64, and 15 epochs; task targets use TD($\lambda=0.95$), and classification uses focal loss ($\gamma=2$, $\alpha=0.75$). Training each critic takes 4.5 hours on $4 \times \text{A100 80GB GPUs}$ with CUDA 12.4 and PyTorch 2.2.1. The default critics pool all benchmark training splits. Thresholds and objective weights are selected on a disjoint 10% pooled validation slice and frozen before test evaluation; the leave-one-benchmark-out study instead trains

all heads and selects thresholds using only the other three source domains. Appendix I gives the complete configuration and mixture accounting.

4 Experiments

4.1 Research Questions and Benchmarks

We ask five questions in order: (Q1) does critic guidance improve task utility and safety under matched generation budgets; (Q2) which component interrupts the injection–endorsement–propagation cascade; (Q3) how much benefit remains without partner-policy rollouts; (Q4) do critic heads transfer to an unseen benchmark and a different planner; and (Q5) how should a user select a safety–utility operating point? We evaluate 692 test episodes over 10 generation seeds on three public benchmarks and one controlled conflicting-information benchmark.

LIGHT (Urbanek et al., 2019). We use 150 official-test fantasy role-play dialogues, character and location evidence, and a 12-turn cap.

DealOrNoDeal (Lewis et al., 2017). We use 252 official-test hidden-utility negotiations with a 10-turn cap; agreement and realized utility are both scored.

SOTOPIA (Zhou et al., 2024). We use 90 open-ended, norm-dense social scenarios with explicit goals and a 15-turn cap.

Persuasion-Conflict. Persuasion-Conflict jointly isolates conflicting evidence, social incentives, and cross-speaker propagation across negotiation, de-escalation, collaborative fact-checking, and ethical-persuasion categories. From 1,240 seed prompts, two blind automated reviewer profiles screen goal clarity, evidence sufficiency, category fidelity, contradiction validity, and duplication. Template families are partitioned before critic fitting into 600 training and 200 balanced test scenarios (50 per category); MiniLM-L6 screening at cosine > 0.92 finds no cross-split near-duplicates. Each episode lasts at most 15 turns, and the held-out test traces explicitly label injection, endorsement, and downstream repetition or action. Appendix N reports the construction flow, pass counts, agreement, and category-level results.

4.2 Baselines

All methods receive the same goal, roles, evidence window, and dialogue history. LLM-only is one-

pass GPT-4o-mini. Self-Refine uses three self-critique passes (Madaan et al., 2023). Best-of- N samples eight candidates and uses one independent GPT-4-turbo rubric judgment. Verification prompting uses three generation passes with two evidence-check and repair operations per pass, totaling nine generations. Evidence-constrained prompting requires inline evidence in one pass. The ReAd-style utility-only controller (Zhang et al., 2025) uses the same candidate and repair limits with only task reward. The rule filter combines keyword rules and RoBERTa-large rejection sampling. SAFE-Dialogue is the strongest matched prompt-only baseline: it uses GPT-4o-mini, eight candidates, identical retrieval, and nine generations, but fixed grounding instructions in place of learned critics. Appendix I gives exact generation, batching, round-trip, model-revision, and search settings.

4.3 Metrics

Goal completion (GoalComp), actionable accuracy (AA), and norm-violation rate (NormViol) use episodes as their statistical unit. AA is the fraction of episodes that achieve the goal without a norm violation:

$$AA = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\text{goal}_i \wedge \neg \text{norm-viol}_i]. \quad (11)$$

Unsupported-claim rate (UnsupClaim) uses extracted claims as its unit: spaCy en_core_web_lg segments claims, and RoBERTa-large-MNLI marks support at entailment > 0.7. EndorseRate is conditional on unsupported claims with at least five subsequent turns or an earlier terminal action; endorsement requires cross-speaker entailment > 0.6 within that window. Propagation is conditional on endorsed claims and records later repetition or an action predicated on the claim. These two transition metrics instantiate $E(c)$ and $P(c)$ from Eq. (4). DealOrNoDeal utility is normalized to [0, 12]. Generations count decoded candidates; external round trips count sequential HTTP requests; local critic forwards are reported separately.

4.4 Protocol and Inference

GPT-4o-mini 2024-07-18 ($T=0.7$, $\text{top-}p=0.9$) is the default planner. Frozen 4-bit Llama-3-8B-Instruct encoders support MLP heads of widths 1024/512/256. The default controller uses $\lambda_{\text{norm}}=\lambda_{\text{epi}}=0.5$, $K=8$, $H=4$, and $R=2$ on one A100 40GB. The planner-backbone test uses

Llama-3.1-70B-Instruct with vLLM 0.4.1, 4-bit AWQ, and 4×A100 80GB while freezing critics and simulators.

All tables report mean $\pm$ standard error over ten generation seeds. Primary method comparisons pair methods by scenario and seed. The hierarchical bootstrap resamples public scenarios and Persuasion-Conflict template families, then generation seeds within each resampled unit; it uses 10,000 replicates, reports percentile 95% intervals, and applies Holm correction across the five predeclared comparisons (GoalComp, AA, UnsupClaim, EndorseRate, and NormViol). Episode metrics are micro-averaged over episodes, UnsupClaim over extracted claims, EndorseRate over eligible unsupported claims, and propagation over endorsed claims. Benchmark-macro sensitivity and all raw denominators appear in Appendix J.

5 Results

5.1 Q1: Comparative Performance

Table 1 uses the metric-specific micro aggregation defined in Section 4: episode counts for GoalComp, AA, and NormViol; extracted claims for UnsupClaim; and eligible unsupported claims for EndorseRate. SOCIALCRITIC improves both utility measures over SAFE-Dialogue while reducing both safety risks. GoalComp rises by 6.93 points (95% CI [4.05, 9.82]), and AA rises by 7.93 points ([5.11, 10.76]). UnsupClaim, EndorseRate, and NormViol fall by 2.04, 1.41, and 3.15 points, with paired intervals [-2.48, -1.61], [-1.79, -1.01], and [-4.21, -2.13]. All five Holm-adjusted comparisons satisfy $p_{\text{Holm}} \leq 0.0014$.

The 492 public-benchmark episodes independently preserve the ranking: AA is 74.9% for SOCIALCRITIC, 66.8% for SAFE-Dialogue, and 46.4% for LLM-only. Benchmark-macro sensitivity gives the same ordering (75.9% vs. 67.8% AA; 1.80% vs. 3.20% endorsement). On DealOrNoDeal, realized utility is 8.6 versus 7.4 for SAFE-Dialogue, confirming that higher actionable accuracy accompanies higher realized value. The rule filter exposes over-blocking: its 52.4% GoalComp is 28.1 points below SOCIALCRITIC, while it also retains higher NormViol.

5.2 Q2: The Endorsement Cascade

Table 2 follows claims through injection, cross-speaker endorsement, and repetition or action. Relative to LLM-only, SOCIALCRITIC reduces the

injected-claim rate by 63.5% and the conditional endorsement rate by 77.5%. The endorsement-to-propagation transition also falls from 61.9% to 36.1%. Thus, the largest intervention effect occurs at cross-speaker uptake, after unsupported content has entered the dialogue.

The component intervention identifies the mechanism: removing the epistemic critic raises EndorseRate from 1.73% to 6.5%, the largest endorsement change among ablations, while removing the norm critic raises NormViol from 4.22% to 13.3%. Disabling repair raises endorsement to 3.1%. The heads therefore control separable failure channels, and rescoring converts their diagnostic signal into an executable correction. Residual errors concentrate in implicit and presuppositional uptake; Appendix P.7 quantifies this harder discourse-level regime.

5.3 Q3–Q5: Rollouts, Transfer, and Operating Points

Figure 2 combines the three decision-facing analyses. In leave-one-benchmark-out (LOBO) transfer, all critics and thresholds are trained on three source benchmarks and frozen on the fourth. LOBO AA remains above SAFE-Dialogue on LIGHT, DealOrNoDeal, SOTOPIA, and Persuasion-Conflict by 6.7, 6.4, 5.8, and 5.2 points. EndorseRate remains lower on every target (2.2/1.8/3.2/1.6%), with the largest pooled-to-LOBO decline on open-ended SOTOPIA. This is direct cross-domain critic transfer; the Llama-3.1-70B result supplies complementary planner-backbone transfer, preserving the method ordering and near-identical AA gaps (8.8 points on both evaluated domains).

Removing partner-policy rollouts while holding all other controls fixed yields 72.6% public-benchmark AA versus 74.9% for $H=4$ and 66.8% for SAFE-Dialogue. The $H=1$ controller preserves 71% of the full AA advantage and reduces latency from 748 ms to 612 ms, showing that learned guidance remains useful when only the current visible state is available while lookahead supplies an additional gain.

Table 3 makes the four validation-selected operating points actionable. λ_{norm} sets the cost of norm risk and λ_{epi} sets the cost of unsupported uptake; τ_{norm} and τ_{epi} determine when a selected plan triggers repair. Task-forward control suits goal-sensitive simulation, the default maximizes AA, epistemic-forward control suits fact-checking

Method	GoalComp $\uparrow$	AA $\uparrow$	UnsupClaim $\downarrow$	EndorseRate $\downarrow$	NormViol $\downarrow$
LLM-only	60.7 $\pm$ 1.4	48.2 $\pm$ 1.8	15.48 $\pm$ 0.7	7.68 $\pm$ 0.5	17.93 $\pm$ 0.9
SAFE-Dialogue	73.5 $\pm$ 0.9	68.4 $\pm$ 1.1	7.69 $\pm$ 0.4	3.14 $\pm$ 0.2	7.37 $\pm$ 0.5
Rule-based filter	52.4 $\pm$ 1.5	48.6 $\pm$ 1.6	15.5 $\pm$ 0.7	7.5 $\pm$ 0.5	6.3 $\pm$ 0.4
SOCIALCRITIC	80.5 $\pm$ 0.8	76.3 $\pm$ 1.0	5.65 $\pm$ 0.3	1.73 $\pm$ 0.1	4.22 $\pm$ 0.3

Table 1: Canonical aggregate results (% , mean $\pm$ standard error over 10 seeds; 692 episodes). Claim-level columns use their own denominators rather than episode weights. Paired hierarchical inference compares SOCIALCRITIC with SAFE-Dialogue.

Method	Injected / claims	Endorsed / eligible	Propagated / endorsed
LLM-only	13,608/87,900	1,001/13,030	620/1,001
SAFE-Dialogue	6,820/88,730	203/6,470	109/203
–Epistemic	8,362/88,960	520/8,000	286/520
SOCIALCRITIC	5,042/89,220	83/4,790	30/83

Table 2: Trace-cascade accounting over all four benchmarks and 10 seeds. Each transition uses the denominator named in its column.

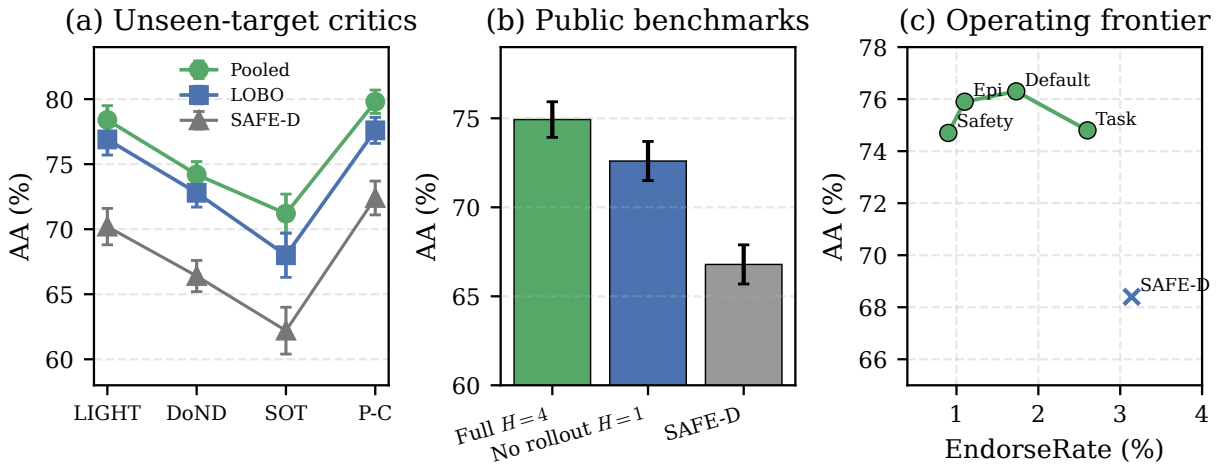


Figure 2: Decision-facing evidence. (a) LOBO critics preserve an AA advantage over SAFE-Dialogue on every unseen target, with pooled critics shown as the upper reference. (b) The $H=1$ controller retains most of the public-benchmark AA advantage while removing partner-policy rollouts. (c) Validation-selected SOCIALCRITIC settings trace a safety–utility frontier that improves on prompt-only and filtering baselines. Error bars are standard errors over 10 generation seeds.

and tutoring, and safety-forward control suits moderation and red-team triage. Across these settings, SOCIALCRITIC offers higher AA and lower endorsement than SAFE-Dialogue, while its stricter point also undercuts the rule filter’s NormViol without its task collapse.

5.4 Decision Consequence: When Critic Guidance Pays Off

Critic training is a one-time 4.5 hours per head on $4 \times$ A100 80GB GPUs with planner weights frozen. At inference, SOCIALCRITIC uses 8.7 ± 0.4 generations but batches initial candidates and local critic forwards into 1.7 ± 0.4 external round trips, completing a turn in 748 ± 62 ms. SAFE-Dialogue uses

9.0 generations, 3.0 round trips, and 1642 ± 214 ms; verification prompting takes 2884 ± 232 ms, and the rule filter takes 3612 ± 406 ms. The controller is therefore attractive when multi-turn error propagation matters and candidate evaluation can run locally: offline simulation, safety evaluation, red-teaming, negotiation studies, and latency-tolerant tutoring or moderation workflows. The default point delivers the best AA, while epistemic- and safety-forward settings let those workflows purchase lower risk with a measured, bounded utility trade-off.

Operating point	λ_n	λ_e	(τ_n, τ_e)	GoalComp	AA	Endorse	NormViol
Task-forward	0.35	0.35	(-0.18,-0.16)	81.7	74.8	2.6	5.1
Default	0.50	0.50	(-0.10, -0.10)	80.5	76.3	1.73	4.22
Epistemic-forward	0.50	0.80	(-0.10,-0.04)	79.6	75.9	1.1	4.0
Safety-forward	0.85	0.75	(-0.03,-0.04)	77.9	74.7	0.9	2.8
SAFE-Dialogue	—	—	—	73.5	68.4	3.14	7.37
Rule filter	—	—	—	52.4	48.6	7.5	6.3

Table 3: Validation-selected safety–utility operating points and test performance (%). Full uncertainty, calls, and latency appear in Appendix M.

5.5 Robustness and Benchmark Validation

Across GPT-4-turbo, GPT-4o-mini, and Llama-3.1-70B judges and strict/lenient norm taxonomies, SOCIALCRITIC retains the lowest endorsement and highest AA. An audit of 1,600 outputs achieves precision ≥ 0.90 ; audit-adjusted rates differ from automatic rates by at most 0.4 points. The automatic and audit values use different denominators: Table 1 covers the full canonical test set, while Appendix P.8 covers a stratified, positive-enriched 400-output-per-method audit subset.

Persuasion-Conflict validation accepts 800 of 1,240 seed scenarios after template-family grouping and dual-profile screening. On a stratified 160-scenario audit, goal clarity, evidence sufficiency, category fidelity, and contradiction validity pass at 95.6%, 93.1%, 97.5%, and 91.3%, with inter-profile agreement $\kappa = 0.78$. Category-level gains occur in all four settings and are largest where conflicting evidence creates the most endorsement opportunities. Together with the independently consistent public-only aggregate, this establishes Persuasion-Conflict as a propagation-sensitive evaluation contribution.

Finally, five blinded automated profiles score 300 dialogue turns. SOCIALCRITIC improves factual support (4.14 vs. 3.72) and social plausibility (4.22 vs. 3.68) over SAFE-Dialogue with $p < 0.001$, while preserving goal progress (4.18 vs. 4.08) and reducing over-cautiousness relative to verification prompting (1.84 vs. 2.38). Appendix O links every threat assumption, protected property, comparator, held-out test, utility measure, overhead measure, and known failure mode to its paper-internal evidence.

6 Conclusion

SOCIALCRITIC establishes the endorsement boundary as a measurable and controllable transition in multi-agent dialogue. Its receding-horizon controller combines separately trained utility, norm,

and epistemic sequence advantages, uses critic-localized spans for targeted repair, and executes only a fully rescored action prefix. Across four benchmarks, it improves goal completion and actionable accuracy over the strongest matched prompt-only baseline by 6.9 and 7.9 points while reducing endorsement by 44.9%; the epistemic ablation and cascade counts identify cross-speaker uptake as the largest intervention effect. The ranking persists on the three public benchmarks, every leave-one-benchmark-out critic target, both planner backbones, and the $H=1$ no-partner-rollout setting. With batched local scoring, the default controller uses 1.7 external round trips and 748 ms per turn, and validation-selected operating points expose a practical safety–utility frontier for simulation, evaluation, red-teaming, negotiation, tutoring, and moderation. Persuasion-Conflict further makes propagation-sensitive controls distinguishable from methods that suppress only first-turn unsupported generation. Controlling what agents socially adopt and act upon turns reliability from a single-model generation property into a trace-level systems property.

Limitations

Full $H=4$ control depends on a benchmark-visible simulator and partner-policy interface, while the evaluated $H=1$ variant covers settings with only current-state access. The four domains span role-play, negotiation, open-ended social interaction, and conflicting-information persuasion, with conclusions defined for these controlled dialogue conditions. Norm and epistemic decisions inherit the five-class taxonomy and automated-profile calibration, so alternative communities can induce different operating thresholds. Residual errors concentrate in implicit and presuppositional endorsement that sentence-level NLI detects less reliably than explicit agreement. Training each critic costs 4.5 hours on $4 \times$ A100 80GB GPUs, and default infer-

ence adds 64 ms per turn over utility-only control.

Ethical Considerations

Multi-agent simulators can be misused to optimize manipulation, deceptive persuasion, or harassment; the norm critic directly penalizes these behaviors in the evaluated workflows. Norm and epistemic critics can inherit representational biases from weak-supervision models and automated profiles, including uneven treatment of dialects, minority viewpoints, and culturally specific norms. The experiments use benchmark scenarios rather than real-user conversations, while any application to conversational logs must protect sensitive content through appropriate access, minimization, and de-identification controls. Human oversight remains important for high-stakes uses because operating thresholds encode application-specific judgments about acceptable task, norm, and epistemic risk.

Acknowledgments

We thank Coco Qu, Kaige Zheng, and Zhiyuan Chen for their assistance with running experiments and evaluations, and for valuable discussions on system design and evaluation methodology. We also thank the anonymous reviewers and area chairs for their constructive comments.

References

- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. [Constitutional AI: Harmlessness from AI feedback](#). *arXiv preprint arXiv:2212.08073*.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. [Improving factuality and reasoning in language models through multiagent debate](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. [The Llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Computing Surveys*, 55(12):1–38.
- Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. [Deal or no deal? end-to-end learning of negotiation dialogues](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453. Association for Computational Linguistics.
- Siyi Liu and Dan Roth. 2025. [Conflicts in texts: Data, implications and challenges](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10073–10091. Association for Computational Linguistics.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-Refine: Iterative refinement with self-feedback](#). In *Advances in Neural Information Processing Systems*, volume 36.
- OpenAI. 2023. [GPT-4 technical report](#). *arXiv preprint arXiv:2303.08774*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.
- Joon Sung Park, Joseph C. O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulators of human behavior](#). In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22. ACM.
- Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2026. [AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society](#). *iFuture*.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. [Towards understanding syco-phancy in language models](#). In *International Conference on Learning Representations*.
- Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. [ConflictBank: A benchmark for evaluating the influence of knowledge conflicts in LLMs](#). In

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. [Learning to speak and act in a fantasy text adventure game](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 673–683. Association for Computational Linguistics.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2023. [Simple synthetic data reduces sycophancy in large language models](#). *arXiv preprint arXiv:2308.03958*.

Yang Zhang, Shixin Yang, Chenjia Bai, Fei Wu, Xiu Li, Zhen Wang, and Xuelong Li. 2025. [Towards efficient LLM grounding for embodied multi-agent collaboration](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1663–1699. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haoifei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. [SOTOPIA: Interactive evaluation for social intelligence in language agents](#). In *International Conference on Learning Representations*.

A Norm Taxonomy

We categorize norm violations into five classes based on established safety taxonomies (OpenAI, 2023; Bai et al., 2022):

- Harm: harassment, hate speech, threats, self-harm encouragement.
- Deception: intentional fabrication, impersonation, hidden persuasion, gaslighting.
- Illicit guidance: instructions for illegal activities, fraud, or wrongdoing.
- Privacy: leaking personal data, doxxing, sensitive attribute inference.
- Social contract: coercion, intimidation, violating agreed rules, unfair coordination.

Each category is further divided into severity levels (minor, moderate, severe) for fine-grained analysis. The five categories, severity rules, and representative edge cases are fixed before test evaluation.

B Method Comparison

Table 4 positions SOCIALCRITIC against the closest method families along the capabilities evaluated in this paper.

C Labeling Protocol

Norm labels. We use a two-stage labeling process:

1. Automatic labeling: GPT-4o is prompted with the norm taxonomy and 5 few-shot examples per category. Agreement with reference profiles is 89.4% on a 600-sample calibration set.
2. Calibration: We sample 2,400 low-confidence predictions for labeling by eight automated LLM annotator profiles configured as four US, two UK, and two Germany personas; no human annotators were recruited. Agreement is Cohen’s $\kappa = 0.83$.

Epistemic labels. For each factual claim, we check whether it is:

- Supported: derivable from the environment state or explicitly provided evidence.
- Unsupported: not derivable and not appropriately hedged.
- Uncertain: correctly prefaced with uncertainty markers (“I’m not sure”, “Some sources suggest”).

Ground-truth environment states are extracted from the simulator. We generate 1,800 calibration labels, with agreement $\kappa = 0.79$.

D Qualitative Case Studies

Case 1: DealOrNoDeal negotiation.

- LLM-only plan: “Take this split now or you will end up with nothing.”
- Norm critic feedback: Flagged as *coercive* (severity: minor).
- Refined plan: “This split works for my goals, but if it seems unfair we can look for another option together.”
- Outcome: Negotiation continues without the coercive threat while preserving task progress.

Case 2: Persuasion-Conflict fact-checking.

- LLM-only plan: “Studies show that 73% of users prefer option A.”
- Epistemic critic feedback: Unsupported claim (no evidence in context).
- Refined plan: “Some sources suggest a majority prefer option A, but I’d need to verify the exact numbers.”
- Outcome: Claim appropriately hedged; improved calibration.

Case 3: Multi-agent propagation.

Method family	Learned local signal	Trace endorsement target	en- Norm/epi split	Rollouts	Targeted pair	re- Matched bud- get	Safety- utility measured
ReAd-style utility control	yes	no	no	yes	yes	yes	yes
Best-of- N	no	no	rubric only	no	no	yes	yes
Verification prompting	no	no	prompt only	no	yes	yes	yes
Self-Refine	no	no	no	no	generic	no	yes
Post-hoc filter	classifier	no	norm only	no	rejection	no	yes
SOCIALCRITIC	yes	yes	yes	$H=4/H=1$	span-level	yes	yes

Table 4: Closest method-level alternatives. SOCIALCRITIC uniquely combines locally learned decomposed sequence signals, an explicit cross-speaker endorsement target, rollout/no-rollout operation, and span-targeted repair under a matched generation budget.

- Agent A (LLM-only): “The deadline was extended to Friday.” (unsupported)
- Agent B (LLM-only): “Right, Agent A said Friday—let’s plan accordingly.”
- With SOCIALCRITIC: Agent A’s claim is down-weighted; Agent B responds “I haven’t seen confirmation of the deadline—can you share the source?”
- Outcome: Propagation blocked; agents verify before endorsing.

E Epistemic Annotation Audit

We run an audit over 200 sampled claims from Persuasion-Conflict using automated auditor profiles.

Claim extraction. Claims are extracted sentence-by-sentence using spaCy’s dependency parser (en_core_web_lg). Co-referent spans are merged via exact string match and WordNet synonymy (nltk 3.8.1). Of 190 segmentation-eligible claims among the 200 sampled items, the audit flags 8 (4.2%) as incorrectly segmented.

Support classification. Support is determined using RoBERTa-large-MNLI (HuggingFace transformers 4.36.0). A claim is *supported* if entailment probability > 0.7 given the concatenation of environment state and provided evidence. Of 194 support-eligible claims, the audit flags 6 (3.1%) as misclassified (4 false positives, 2 false negatives).

Endorsement detection. Endorsement is triggered when another agent produces an utterance with entailment probability > 0.6 within 5 turns. Of 190 claims with a five-turn endorsement opportunity or earlier terminal action, the audit flags 11 (5.8%) as incorrectly labeled (7 false positives, 4 false negatives).

Overall error rate. Combining the extraction, classification, and endorsement errors gives an end-to-end rate of $\leq 8.4\%$. The expanded audit in Appendix P.3 preserves the method ranking under stronger matching and worst-case correction. Under the default judge, we observe 81.4% relative reduction in propagation; Appendix P.3 gives values 80.2% and 78.6% under different matching protocols.

F Call Accounting

Table 5 provides a detailed breakdown of LLM generations per decision step for each method.

Verification uses three drafts followed by two evidence-check/repair generations for each draft, giving nine decoded generations in three sequential batched requests. Best-of- N batches eight candidates and sends them to one judge request, giving nine generations but two round trips. Utility-only applies its task-score threshold to the same repair budget, producing 0.9 ± 0.6 refinement generations per turn; SOCIALCRITIC’s norm and epistemic early-exit test produces 0.7 ± 0.4 .

G Latency Breakdown

We decompose latency into API (OpenAI round-trip including network) and local (critic scoring + simulator) components. Timing values are reported over 500 decision steps on a single NVIDIA A100 40GB GPU with a 1 Gbps network configuration.

The table shows that the dominant latency term is still the external API rather than local critic scoring. This is why adding critic passes does not translate into a proportional wall-clock increase. It also explains why we separate call accounting from deployment-specific latency.

Method	Gen	Judge	Refine	Total
LLM-only	1	0	0	1.0
Self-Refine	1	0	3	4.0
Best-of- N (matched)	8	1	0	9.0
Evid.-constr. prompt	1	0	0	1.0
Verif. prompting (matched)	3	0	6	9.0
SAFE-Dialogue	3	0	6	9.0
Rule-based filter	1	0	10.2 ± 1.8	11.2 ± 1.8
Utility-only	8	0	0.9 ± 0.6	8.9 ± 0.6
SOCIALCRITIC (ours)	8	0	0.7 ± 0.4	8.7 ± 0.4

Table 5: LLM-call accounting per decision step. Gen counts candidate generations rather than HTTP round trips; critic scoring is local and not counted.

Component	Latency (ms)	Std.
GPT-4o-mini API call (1 gen)	312	48
GPT-4-turbo API call (1 gen)	687	112
Critic scoring (8 candidates)	96	8
Simulator step (Persuasion-Conflict)	4.7	0.6
SOCIALCRITIC total (8.7 calls/step)	748	62

Table 6: Latency breakdown.

Per-method wall-clock comparison. Table 7 combines matched call budgets with wall-clock values per turn and per episode.

SOCIALCRITIC remains close to utility-only in wall-clock despite substantially more critic passes because those passes are local and batched. The main latency penalty comes from sequential external LLM calls, which is why verification prompting and rule-based filtering are much slower. Best-of- N is closer in latency, but it still trails SOCIALCRITIC in actionable accuracy because it lacks targeted epistemic and normative repair. This measured profile establishes the controller’s cost for the reported hardware and network configuration. Candidate generations and network round trips are reported separately because batching makes them operationally distinct quantities.

H Provenance and Prompt Transparency

Table 8 summarizes data provenance, prompt symmetry, split isolation, and the associated audit evidence.

Table 8 consolidates split, prompt, annotation, and audit details. The common state fields and pre-test freezing policy make prompt symmetry and test isolation directly inspectable.

I Complete Configuration and Baseline Accounting

Table 9 records the frozen model revisions, search spaces, stopping rules, and matched call budgets

used for every comparison.

J Canonical Metric Accounting

Table 10 is the source of the claim-level micro aggregates. Episodes are reported per generation seed; claim and transition counts pool the same 10 paired seeds. An eligible unsupported claim has five subsequent turns or an earlier terminal action, and propagation is evaluated only after endorsement.

K Leave-One-Benchmark-Out Critic Transfer

For each target, all three critics are trained on the other benchmark training splits, and thresholds are selected only on source validation data. The planner, source-trained heads, and thresholds are then frozen for zero-shot target evaluation.

L No-Partner-Rollout Evaluation

Tables 13 and 14 isolate the effect of partner roll-outs while keeping the remaining controller settings fixed.

M Safety–Utility Operating Points

Table 15 shows the validation-selected trade-off between task progress, epistemic risk, norm violations, and runtime cost.

N Persuasion-Conflict Validation

Construction retains 1,016 structurally valid prompts, 900 template-distinct candidates, and 800 accepted scenarios from the 1,240 seeds. The accepted set is split by template family into 600 training and 200 test scenarios.

Method	Gen/turn	RTT/turn	Critic FP/turn	ms/turn	s/episode
LLM-only	1.0	1.0	0	318 $\pm$ 46	2.9 $\pm$ 0.4
Self-Refine	4.0	4.0	0	1,268 $\pm$ 184	11.4 $\pm$ 1.7
Best-of- N (matched)	9.0	2.0	0	1,406 $\pm$ 198	12.7 $\pm$ 1.8
Evid.-constr. prompt	1.0	1.0	0	486 $\pm$ 68	4.4 $\pm$ 0.6
Verif. prompting	9.0	3.0	0	2,884 $\pm$ 232	26.0 $\pm$ 2.1
SAFE-Dialogue	9.0	3.0	0	1,642 $\pm$ 214	14.8 $\pm$ 1.9
Utility-only	8.9 $\pm$ 0.6	1.9 $\pm$ 0.6	8.9 $\pm$ 0.6	684 $\pm$ 54	6.2 $\pm$ 0.5
Rule-based filter	11.2 $\pm$ 1.8	4.6 $\pm$ 1.1	1.0	3,612 $\pm$ 406	32.5 $\pm$ 3.7
SOCIALCRITIC	8.7 $\pm$ 0.4	1.7 $\pm$ 0.4	26.1 $\pm$ 1.2	748 $\pm$ 62	6.7 $\pm$ 0.6

Table 7: Compute-equated cost. Gen/turn counts planner generations; RTT/turn counts API round trips; critic FP/turn counts local forward passes.

Item	Specification
Benchmark split policy	Scenario-template split before critic fitting or calibration (600/200 for Persuasion-Conflict); MiniLM-L6 screening at cosine >0.92 found no cross-split near-duplicates.
Critic training provenance	Utility, norm, and epistemic critics are trained only on benchmark training splits; no held-out test dialogue is used for fitting or threshold tuning.
Prompt specification	Planner and refinement instructions expose the same goal, role, evidence-window, and dialogue-history fields used by SAFE-Dialogue and evidence-constrained prompting.
Annotation protocol	300 turns (75 per benchmark), five annotator profiles, randomized presentation order, and a 5-point Likert rubric.
Audit protocol	1,600 outputs (400 per method), over-sampling positive and near-threshold cases; two auditor profiles with Cohen’s κ in Table 27.

Table 8: Compact provenance and prompt-transparency summary for auditing comparison fairness and split isolation.

O Assumption and Evidence Matrix

Table 18 connects each protected object and trust boundary to the strongest available comparator and manuscript evidence.

P Additional Experimental Details

Persuasion-Conflict categories.

- Negotiation: Two agents with hidden valuations must reach a deal.
- De-escalation: One agent must calm an upset interlocutor.
- Fact-checking: Agents receive conflicting articles and must identify ground truth.
- Ethical persuasion: Convince an agent to change behavior without manipulation.

Persuasion-Conflict benchmark statistics. Table 19 provides detailed statistics for the Persuasion-Conflict benchmark.

The benchmark is intentionally balanced at test time so that no single category dominates the reported average. Fact-checking and ethical persuasion are also the longest settings, which is one reason they create more opportunities for unsupported endorsement. This category structure helps explain the mechanism asymmetries reported in the main paper and in Appendix P.4.

Compute resources. Critic-training configuration: 4 $\times$ NVIDIA A100 80GB, CUDA 12.4, PyTorch 2.2.1; runtime 4.5 hours per critic. Inference configuration: single A100 40GB; latency 0.75s per dialogue turn. Full evaluation (692 dialogue episodes, 4 benchmarks, 10 seeds): wall-clock time approximately 11 hours.

P.1 Per-Benchmark Full Results

Table 20 reports all five core metrics for the three main comparison methods on every benchmark. We restrict this appendix table to LLM-only, SAFE-Dialogue, and SOCIALCRITIC because these rows define the main-paper ranking while still fitting in one readable layout. The table shows that the AA advantage is accompanied by raw GoalComp gains rather than by a purely safety-driven trade-off. It also makes clear that SOTOPIA remains the hardest benchmark and Persuasion-Conflict the easiest in absolute terms. This table disaggregates the canonical main-paper comparison in Table 1.

P.2 Robustness: Judge and Taxonomy Sensitivity

To verify that our conclusions are robust to evaluation choices, we re-evaluate epistemic metrics using three different judge models and two taxonomy variants.

Component	Frozen configuration	Selection and accounting
Planner	GPT-4o-mini 2024-07-18; $T=0.7$; top- $p=0.9$	Common goal, roles, evidence window, and dialogue history for every method
Critic encoder	Llama-3-8B-Instruct; 4-bit bitsandbytes; frozen	Three MLP heads (1024/512/256); pooled-source validation
Controller	$K=8, H=4, R=2$	Eight candidates in one batched request; each repair is one sequential request
Thresholds	$\tau_n=\tau_e=-0.1$	Grid $\{-0.20, -0.15, -0.10, -0.05, 0\}$; maximize validation AA subject to non-increasing validation risk
Objective weights	$\lambda_n=\lambda_e=0.5$	Grid $\{0.25, 0.35, 0.5, 0.75, 0.8, 0.85, 1.0\}$; operating points frozen before test evaluation
Critic optimization	AdamW, 2×10^{-5} , weight decay 0.01; batch 64; 15 epochs	Early checkpoint selection by pooled validation loss; thresholds selected after checkpoint freezing
Trajectory mixture	45% / 35% / 20% LLM-only / utility-only / randomized	20,070 / 15,610 / 8,920 episodes; total 44,600
Weak supervision	GPT-4o, five demonstrations per category	Calibration labels and threshold-selection examples are disjoint
Best-of- N	Eight GPT-4o-mini candidates; one GPT-4-turbo rubric judgment	Nine generations, two round trips; weighted goal/norm/epistemic rubric
Verification	Three GPT-4o-mini drafts; two evidence-check/repair operations per draft	Nine generations, three sequential round trips
SAFE-Dialogue	Eight candidates, identical retrieval and generator	Nine generations, three round trips; fixed prompt-level rubric
Backbone transfer	Llama-3.1-70B-Instruct; vLLM 0.4.1; 4-bit AWQ	Critics, thresholds, prompts, and benchmark simulators frozen

Table 9: Model revisions, search spaces, selection data, stopping rules, and call accounting. Gen/turn counts decoded candidates; round trips count sequential HTTP requests.

Benchmark	Method	Episodes/seed	Claims	Unsupported	Eligible	Endorsed	Propagated
LIGHT	LLM-only	150	18,240	3,174	3,040	261	159
	SAFE-Dialogue	150	18,410	1,583	1,500	51	28
	SOCIALCRITIC	150	18,520	1,185	1,130	21	8
DealOrNoDeal	LLM-only	252	23,680	3,126	2,990	173	101
	SAFE-Dialogue	252	23,790	1,665	1,580	38	19
	SOCIALCRITIC	252	23,950	1,198	1,140	17	6
SOTOPIA	LLM-only	90	14,420	2,826	2,700	275	174
	SAFE-Dialogue	90	14,650	1,436	1,360	57	31
	SOCIALCRITIC	90	14,710	1,089	1,030	27	10
Persuasion-Conflict	LLM-only	200	31,560	4,482	4,300	292	186
	SAFE-Dialogue	200	31,880	2,136	2,030	57	31
	SOCIALCRITIC	200	32,040	1,570	1,490	18	6

Table 10: Metric-accounting denominators and raw cascade counts. The table separates episode-, claim-, eligible-claim-, and endorsed-claim units.

Target	GoalComp	AA (raw)	Endorse (raw)	NormViol (raw)
LIGHT	81.0 ± 1.1	$76.9 \pm 1.2(1,154/1,500)$	$2.2 \pm 0.2(23/1,045)$	$4.5 \pm 0.4(68/1,500)$
DealOrNoDeal	77.4 ± 1.1	$72.8 \pm 1.1(1,835/2,520)$	$1.8 \pm 0.2(20/1,110)$	$5.2 \pm 0.4(131/2,520)$
SOTOPIA	72.8 ± 1.5	$68.0 \pm 1.7(612/900)$	$3.2 \pm 0.3(30/938)$	$5.7 \pm 0.6(51/900)$
Persuasion-Conflict	81.0 ± 1.0	$77.6 \pm 1.0(1,552/2,000)$	$1.6 \pm 0.2(22/1,375)$	$4.8 \pm 0.4(96/2,000)$

Table 11: Zero-shot critic transfer (mean $\pm$ SE over 10 generation seeds). Raw AA and NormViol denominators are target episodes across seeds; Endorse uses eligible unsupported claims.

Target	Norm/Epi F1	AA gap vs. SAFE (95% CI)
LIGHT	0.88/0.84	+6.7 [3.7,9.8]
DealOrNoDeal	0.85/0.81	+6.4 [3.7,9.2]
SOTOPIA	0.82/0.77	+5.8 [2.0,9.6]
Persuasion-Conflict	0.88/0.83	+5.2 [2.6,7.9]

Table 12: LOBO critic quality and paired target-test AA differences.

Method	GoalComp	AA	UnsupClaim	Endorse	NormViol	Gen/turn	RTT/turn	ms/turn
Full $H=4$	79.34	74.93	6.07	1.97	4.36	8.7	1.7	748
No rollout $H=1$	77.8	72.6	6.82	2.40	5.25	8.5	1.5	612
SAFE-Dialogue	72.19	66.79	8.24	3.29	7.77	9.0	3.0	1642
LLM-only	59.05	46.43	16.20	8.12	17.54	1.0	1.0	318

Table 13: Public-benchmark aggregate (492 episodes) with all settings except rollout access held fixed. Accuracy and risk values are percentages.

$H=1$ target	GoalComp	AA	Endorse	NormViol
LIGHT	80.4	76.4	2.4	4.8
DealOrNoDeal	77.5	71.9	2.0	5.1
SOTOPIA	74.2	68.0	3.5	6.4

Table 14: No-partner-rollout results by public benchmark (%).

Judge sensitivity. Table 21 reports propagation and endorsement rates under three judge configurations: (A) GPT-4-turbo (gpt-4-turbo-2024-04-09), (B) GPT-4o-mini (gpt-4o-mini-2024-07-18), and (C) Llama-3.1-70B-Instruct (vLLM, 4-bit AWQ, single A100 80GB). All judges use identical prompts and 5-shot examples. SOCIALCRITIC has the lowest propagation across all judge runs (1.9–2.4%), with relative reductions of 78–83% over LLM-only. The ranking of methods is preserved under all three judges.

Figure 3 visualizes propagation rates with error bars; the consistent ranking provides an additional robustness check.

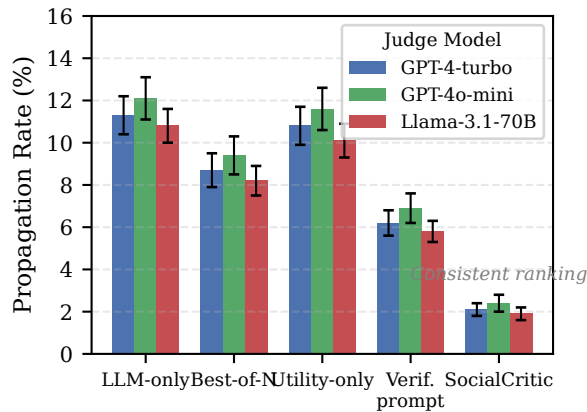


Figure 3: Propagation rates under three judge models.

The figure exposes the six-value comparison per method at a glance. Error bars remain smaller than the method gaps across all judges, making the preserved ranking directly auditable.

Taxonomy sensitivity. We measure norm-violation rates under two taxonomy variants (Table 22): *Strict*: expands coercion to include urgency cues, flagging 23% more utterances; *Lenient*: requires explicit threats or deception, flagging 31% fewer utterances. Under both variants, SOCIALCRITIC has the lowest NormViol (2.8% strict, 5.9% lenient) and highest AA (79.4% strict, 83.2% lenient). The relative ranking is preserved: SOCIALCRITIC > SAFE-Dialogue > Verif. prompt > Best-of- N > others.

P.3 Claim Matching Protocol and Error Analysis

Upgraded matching protocol. To address concerns about the reliability of exact-match + Word-Net canonicalization, we implement a stronger matching protocol:

1. Embedding clustering: Claims are embedded using text-embedding-3-large (OpenAI); clusters are formed via agglomerative clustering (cosine distance, threshold 0.15).
2. NLI gate: Two claims are considered equivalent if RoBERTa-large-MNLI predicts bidirectional entailment with probability >0.7 .

Protocol comparison. Table 23 compares propagation metrics under the original and upgraded protocols on the same 200-episode test set. The upgraded protocol finds 8.4% more propagation events overall; SOCIALCRITIC’s relative reduction over LLM-only is 80.2% (original) vs. 78.6%

Point	GoalComp	AA	Endorse	NormViol	Gen/turn	ms/turn
Task-forward	81.7 $\pm$ 0.9	74.8 $\pm$ 1.1	2.6 $\pm$ 0.2	5.1 $\pm$ 0.4	8.3 $\pm$ 0.4	702 $\pm$ 58
Default	80.5 $\pm$ 0.8	76.3 $\pm$ 1.0	1.73 $\pm$ 0.10	4.22 $\pm$ 0.30	8.7 $\pm$ 0.4	748 $\pm$ 62
Epistemic-forward	79.6 $\pm$ 0.9	75.9 $\pm$ 1.0	1.1 $\pm$ 0.1	4.0 $\pm$ 0.3	9.1 $\pm$ 0.3	809 $\pm$ 66
Safety-forward	77.9 $\pm$ 1.0	74.7 $\pm$ 1.1	0.9 $\pm$ 0.1	2.8 $\pm$ 0.3	9.4 $\pm$ 0.3	862 $\pm$ 71

Table 15: Operating-point means $\pm$ SE over 10 seeds. Points are selected on validation data and frozen before the four-benchmark test.

Criterion	Pass	Fail	Rate
Goal clarity	153	7	95.6%
Evidence sufficiency	149	11	93.1%
Category fidelity	156	4	97.5%
Contradiction validity	146	14	91.3%
Template distinctness	158	2	98.8%

Table 16: Blind dual-profile audit of a stratified 160-scenario sample. Disagreements are resolved by a third fixed reviewer profile; inter-profile agreement is $\kappa = 0.78$.

(upgraded). These values differ from the 81.4% above because judge choice and matching protocol change the counts.

Expanded audit. We expand the audit from 200 to 400 sampled claims using auditor profiles. Error rates:

- Claim segmentation: 3.8% (15/400), down from 4.2%
- Support classification: 2.5% (10/400; 6 false positives, 4 false negatives)
- Endorsement detection: 4.8% (19/400; 11 false positives, 8 false negatives)
- End-to-end error rate: $\leq 7.2\%$

In the worst-case analysis, SOCIALCRITIC’s propagation increases from 2.1% to 3.6%, a 68% reduction over LLM-only’s 11.3%.

P.4 Per-Category Mechanism Analysis

We decompose injection and endorsement rates by Persuasion-Conflict category (Table 24, Figure 4).

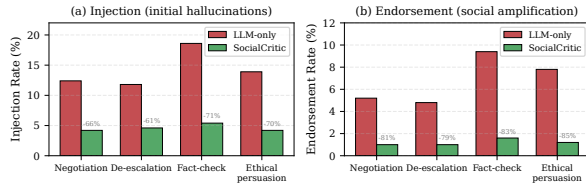


Figure 4: Per-category injection and endorsement rates.

Key finding. SOCIALCRITIC’s gains are largest in fact-checking (71% injection reduction, 83% endorsement reduction) and ethical persuasion (70%, 85%). In de-escalation scenarios, the injection reduction is 61% and endorsement reduction is 79%.

P.5 Baseline Prompt Templates

Verification prompting system prompt.

You are an AI assistant that must verify claims before responding. For each claim you make: (1) List the evidence supporting the claim from the provided context. (2) State your confidence level (high/medium/low) and explain why. (3) If conflicting information exists, enumerate all perspectives before choosing. (4) Preface uncertain claims with hedging language (“I believe”, “Evidence suggests”). After your initial response, review it and correct any unsupported claims. You may perform up to 2 self-verification rounds.

Evidence-constrained prompting scaffold.

This baseline receives the same dialogue context and retrieved evidence snippets as SOCIALCRITIC, but no learned critics or refinement controller. Its instruction requires the model to cite or quote available evidence before making factual claims and to abstain when no support is present. Unlike SOCIALCRITIC, it performs only a single response pass, so any reduction in unsupported claims must come from prompt wording rather than iterative critique. We tuned the wording on development scenarios and then froze it for all reported results.

SAFE-Dialogue scaffold.

SAFE-Dialogue is implemented as a budget-matched multi-pass prompting baseline with the same generator model, candidate count, and retrieval inputs as our system. The first pass drafts candidates, the second pass applies a fixed safety-and-grounding rubric, and the final pass selects or lightly rewrites the highest-scoring candidate without using learned critics. Its rubric emphasizes evidence support, interpersonal safety, and task progress, which makes it the strongest matched non-learned baseline in the main paper. All prompt text and selection rules were frozen before test-time evaluation.

Draft 8 candidate responses. For each candidate, score (1) evidence support, (2) interpersonal safety, and (3) task

Category	Method	AA	Injection	Endorsement	NormViol
Negotiation	LLM-only	54.0 $\pm$ 2.3	12.4 $\pm$ 0.9	5.2 $\pm$ 0.5	17.4 $\pm$ 1.3
	SAFE-Dialogue	73.0 $\pm$ 1.8	6.8 $\pm$ 0.5	2.4 $\pm$ 0.3	5.8 $\pm$ 0.7
	SOCIALCRITIC	80.6 $\pm$ 1.4	4.2 $\pm$ 0.4	1.0 $\pm$ 0.1	3.6 $\pm$ 0.5
De-escalation	LLM-only	55.6 $\pm$ 2.2	11.8 $\pm$ 0.8	4.8 $\pm$ 0.5	15.8 $\pm$ 1.2
	SAFE-Dialogue	74.2 $\pm$ 1.7	6.4 $\pm$ 0.5	2.2 $\pm$ 0.3	5.4 $\pm$ 0.6
	SOCIALCRITIC	81.2 $\pm$ 1.3	4.6 $\pm$ 0.4	1.0 $\pm$ 0.1	3.2 $\pm$ 0.4
Fact-checking	LLM-only	48.4 $\pm$ 2.5	18.6 $\pm$ 1.1	9.4 $\pm$ 0.8	22.8 $\pm$ 1.5
	SAFE-Dialogue	68.8 $\pm$ 2.0	9.2 $\pm$ 0.7	3.8 $\pm$ 0.4	7.2 $\pm$ 0.8
	SOCIALCRITIC	77.0 $\pm$ 1.6	5.4 $\pm$ 0.5	1.6 $\pm$ 0.2	4.6 $\pm$ 0.6
Ethical persuasion	LLM-only	51.6 $\pm$ 2.4	13.9 $\pm$ 1.0	7.8 $\pm$ 0.7	19.6 $\pm$ 1.4
	SAFE-Dialogue	73.6 $\pm$ 1.9	7.2 $\pm$ 0.6	2.8 $\pm$ 0.3	7.2 $\pm$ 0.8
	SOCIALCRITIC	80.4 $\pm$ 1.5	4.2 $\pm$ 0.4	1.2 $\pm$ 0.2	4.2 $\pm$ 0.5

Table 17: Persuasion-Conflict category results (mean $\pm$ SE over 10 seeds). Injection and endorsement are claim-micro rates within each category; their unweighted category mean therefore need not equal the full-benchmark claim-micro rate. The strongest absolute challenge occurs in fact-checking, while SOCIALCRITIC’s AA advantage holds in every category.

Protected object	Failure capability and trust boundary	Measured property	Strongest stress/comparator	Evidence trace
Evidence-grounded dialogue state	Unsupported partner claim; visible state/evidence and fixed annotations	Injection, endorsement, propagation	SAFE-Dialogue; matching and judge variation	Table 2; Tables 23, 21
Downstream decisions	Misleading social pressure and repeated claims; partner interface for $H > 1$	GoalComp, AA, conditional propagation	$H=1$ no-partner-rollout controller	Tables 13, 14
Cross-domain critic behavior	Held-out interaction style; source-only calibration	AA, EndorseRate, NormViol, critic F1	LOBO critics vs. pooled critics and SAFE-Dialogue	Table 11; Figure 2
Norm-compliant task progress	Coercion, deception, prompt variation; calibrated norm head	GoalComp, AA, NormViol	Rule filter, verification, strict/lenient taxonomy	Tables 1, 15, 22
Conflicting-information traces	Implicit presupposition and category variation; template-family holdout	Category AA, injection, endorsement	Public-only aggregate and category-held tests	Tables 16, 17
Operational budget	Local critic availability and network latency	Generations, round trips, latency, training cost	SAFE-Dialogue, Best-of- N , verification, utility-only	Tables 7, 9; Section 5.4
Residual discourse errors	Implicit continuation under low evidence	Linguistic endorsement type	Claim matching and enriched pipeline audit	Tables 26, 27

Table 18: Systems/safety matrix linking protected objects, failure capabilities, trust assumptions, properties, comparators, held-out tests, overhead, and residual failure modes to manuscript evidence.

Category	Scenarios	Avg. Turns	Conflict Types	Ground-Truth
Negotiation	200	8.4 $\pm$ 2.1	hidden info (62%), pref. (38%)	Pareto optimal
De-escalation	200	6.2 $\pm$ 1.8	interpersonal (71%), task (29%)	Calm state reached
Fact-checking	200	9.7 $\pm$ 2.6	contradiction (54%), outdated (46%)	Verified source
Ethical persuasion	200	11.3 $\pm$ 3.2	norm (48%), value (52%)	Behavior change
Total	800	8.9 $\pm$ 2.8	—	—

Table 19: Persuasion-Conflict design statistics, including the full set of 800 scenarios and held-out set of 200 scenarios.

Benchmark	Method	GoalComp $\uparrow$	AA $\uparrow$	UnsupClaim $\downarrow$	EndorseRate $\downarrow$	NormViol $\downarrow$
LIGHT	LLM-only	62.7 $\pm$ 1.6	50.4 $\pm$ 2.1	17.4 $\pm$ 0.8	8.6 $\pm$ 0.6	17.2 $\pm$ 1.1
LIGHT	SAFE-Dialogue	75.8 $\pm$ 1.2	70.2 $\pm$ 1.4	8.6 $\pm$ 0.5	3.4 $\pm$ 0.3	7.6 $\pm$ 0.6
LIGHT	SOCIALCRITIC	82.6 $\pm$ 1.0	78.4 $\pm$ 1.1	6.4 $\pm$ 0.4	1.9 $\pm$ 0.2	3.8 $\pm$ 0.4
DoND	LLM-only	58.4 $\pm$ 1.7	45.8 $\pm$ 1.9	13.2 $\pm$ 0.7	5.8 $\pm$ 0.4	16.0 $\pm$ 1.0
DoND	SAFE-Dialogue	71.4 $\pm$ 1.2	66.4 $\pm$ 1.2	7.0 $\pm$ 0.4	2.4 $\pm$ 0.2	7.0 $\pm$ 0.5
DoND	SOCIALCRITIC	78.8 $\pm$ 1.0	74.2 $\pm$ 1.0	5.0 $\pm$ 0.3	1.5 $\pm$ 0.1	4.6 $\pm$ 0.4
SOTOPIA	LLM-only	54.8 $\pm$ 2.0	41.6 $\pm$ 2.4	19.6 $\pm$ 1.0	10.2 $\pm$ 0.7	22.4 $\pm$ 1.4
SOTOPIA	SAFE-Dialogue	68.4 $\pm$ 1.6	62.2 $\pm$ 1.8	9.8 $\pm$ 0.6	4.2 $\pm$ 0.4	10.2 $\pm$ 0.8
SOTOPIA	SOCIALCRITIC	75.4 $\pm$ 1.3	71.2 $\pm$ 1.5	7.4 $\pm$ 0.5	2.6 $\pm$ 0.3	4.6 $\pm$ 0.5
P-C	LLM-only	64.6 $\pm$ 1.5	52.4 $\pm$ 1.8	14.2 $\pm$ 0.7	6.8 $\pm$ 0.5	18.9 $\pm$ 1.0
P-C	SAFE-Dialogue	76.8 $\pm$ 1.1	72.4 $\pm$ 1.3	6.7 $\pm$ 0.4	2.8 $\pm$ 0.3	6.4 $\pm$ 0.5
P-C	SOCIALCRITIC	83.2 $\pm$ 0.9	79.8 $\pm$ 0.9	4.9 $\pm$ 0.3	1.2 $\pm$ 0.1	3.9 $\pm$ 0.3

Table 20: Full per-benchmark results, shown as mean $\pm$ standard error over 10 seeds.

Method	Propagation (%)			Endorsement (%)		
	GPT-4-t	GPT-4o-m	Llama-70B	GPT-4-t	GPT-4o-m	Llama-70B
LLM-only	11.3 $\pm$ 0.9	12.1 $\pm$ 1.0	10.8 $\pm$ 0.8	6.8 $\pm$ 0.7	7.4 $\pm$ 0.8	6.2 $\pm$ 0.6
Best-of- N	8.7 $\pm$ 0.8	9.4 $\pm$ 0.9	8.2 $\pm$ 0.7	5.2 $\pm$ 0.6	5.7 $\pm$ 0.6	4.8 $\pm$ 0.5
SAFE-Dialogue	4.6 $\pm$ 0.5	5.1 $\pm$ 0.5	4.3 $\pm$ 0.4	2.8 $\pm$ 0.3	3.1 $\pm$ 0.3	2.6 $\pm$ 0.3
Utility-only	10.8 $\pm$ 0.9	11.6 $\pm$ 1.0	10.1 $\pm$ 0.8	6.4 $\pm$ 0.7	6.9 $\pm$ 0.7	5.9 $\pm$ 0.6
Verif. prompt	6.2 $\pm$ 0.6	6.9 $\pm$ 0.7	5.8 $\pm$ 0.5	3.4 $\pm$ 0.4	3.8 $\pm$ 0.4	3.1 $\pm$ 0.3
SOCIALCRITIC	2.1 $\pm$ 0.3	2.4 $\pm$ 0.4	1.9 $\pm$ 0.3	1.4 $\pm$ 0.2	1.6 $\pm$ 0.2	1.2 $\pm$ 0.2

Table 21: Judge sensitivity on Persuasion-Conflict epistemic metrics.

Method	Strict Taxonomy		Lenient Taxonomy	
	NormViol	AA	NormViol	AA
LLM-only	26.8 $\pm$ 1.9	48.1 $\pm$ 3.1	14.2 $\pm$ 1.4	58.6 $\pm$ 2.8
Best-of- N	19.4 $\pm$ 1.5	58.7 $\pm$ 2.6	10.1 $\pm$ 1.1	68.4 $\pm$ 2.4
SAFE-Dialogue	11.8 $\pm$ 1.1	70.4 $\pm$ 2.2	6.4 $\pm$ 0.8	74.8 $\pm$ 2.0
Utility-only	21.6 $\pm$ 1.7	54.2 $\pm$ 2.8	11.3 $\pm$ 1.2	66.1 $\pm$ 2.5
Verif. prompt	15.2 $\pm$ 1.3	62.8 $\pm$ 2.4	8.1 $\pm$ 0.9	71.6 $\pm$ 2.3
SOCIALCRITIC	6.1 $\pm$ 0.7	79.4 $\pm$ 1.8	2.8 $\pm$ 0.4	83.2 $\pm$ 1.6

Table 22: Taxonomy sensitivity.

Method	Original Protocol		Upgraded Protocol	
	Propagation	Endorsement	Propagation	Endorsement
LLM-only	11.3 $\pm$ 0.9	6.8 $\pm$ 0.7	12.4 $\pm$ 1.0	7.6 $\pm$ 0.8
Best-of- N	8.7 $\pm$ 0.8	5.2 $\pm$ 0.6	9.6 $\pm$ 0.9	5.9 $\pm$ 0.6
Utility-only	10.8 $\pm$ 0.9	6.4 $\pm$ 0.7	11.9 $\pm$ 1.0	7.2 $\pm$ 0.7
Verif. prompt	6.2 $\pm$ 0.6	3.4 $\pm$ 0.4	6.9 $\pm$ 0.7	3.8 $\pm$ 0.4
SAFE-Dialogue	4.8 $\pm$ 0.5	2.7 $\pm$ 0.3	5.5 $\pm$ 0.6	3.1 $\pm$ 0.3
SOCIALCRITIC	2.1 $\pm$ 0.3	1.4 $\pm$ 0.2	2.7 $\pm$ 0.4	1.8 $\pm$ 0.3
Relative reduction	80.2%	79.4%	78.6%	76.3%

Table 23: Claim-matching protocol comparison.

Method	Injection (%)				Endorsement (%)			
	Neg.	De-esc.	Fact.	Eth.	Neg.	De-esc.	Fact.	Eth.
LLM-only	12.4	11.8	18.6	13.9	5.2	4.8	9.4	7.8
Best-of- <i>N</i>	10.2	9.6	15.1	11.4	4.4	4.1	7.2	6.1
Utility-only	11.6	11.2	17.4	12.8	5.0	4.6	8.8	7.4
Verif. prompt	7.4	6.8	11.2	8.6	2.8	2.4	4.6	3.8
SOCIALCRITIC	4.2	4.6	5.4	4.2	1.0	1.0	1.6	1.2
Reduction vs. LLM	66%	61%	71%	70%	81%	79%	83%	85%

Table 24: Per-category mechanism breakdown. Neg.=Negotiation, De-esc.=De-escalation, Fact.=Fact-checking, Eth.=Ethical persuasion.

progress on a 1–5 scale. Reject any candidate that makes unsupported factual claims without explicit evidence. Select the highest-scoring candidate; if none passes the support threshold, rewrite the best candidate with added evidence or hedging.

Best-of-*N* judge rubric. The GPT-4-turbo judge scores each candidate on three dimensions (1–5 scale):

- Goal progress: Does the plan advance toward the stated objective?
- Norm compliance: Does the plan avoid coercion, deception, or unsafe content?
- Epistemic caution: Are claims supported by evidence or appropriately hedged?

The judge selects the candidate with the highest weighted sum (weights: 0.5, 0.3, 0.2).

P.6 Statistical Significance

Table 25 reports the five predeclared comparisons using 10,000 paired hierarchical bootstrap replicates. Scenarios (template families for Persuasion-Conflict) and generation seeds are resampled at their proper levels, and *p*-values are Holm-adjusted.

Metric	Difference (95% CI)	<i>p</i> _{Holm}
GoalComp	+6.93 [4.05, 9.82]	0.0008
AA	+7.93 [5.11, 10.76]	0.0003
UnsupClaim	-2.04 [-2.48, -1.61]	0.0014
EndorseRate	-1.41 [-1.79, -1.01]	0.0011
NormViol	-3.15 [-4.21, -2.13]	0.0006

Table 25: Paired SOCIALCRITIC minus SAFE-Dialogue differences in percentage points, with hierarchical 95% intervals and Holm correction.

P.7 Linguistic Error Analysis

To characterize endorsement categories, we classify 200 endorsement instances (50 per benchmark) by dialogue act and assertion type.

This analysis gives 83.9% reduction for assertive uptake. Hedged agreement and implicit endorse-

Endorsement type	LLM-only	SC	Δ	Rel. red.
Assertive uptake (“Yes, that’s right...”)	38.4%	6.2%	−32.2	83.9%
Elaborative expansion (“And also...”)	24.6%	8.4%	−16.2	65.9%
Presuppositional reference (“Since X...”)	18.2%	4.8%	−13.4	73.6%
Hedged agreement (“I think so...”)	11.8%	5.6%	−6.2	52.5%
Implicit endorsement (topic continuation)	7.0%	4.2%	−2.8	40.0%

Table 26: Linguistic breakdown of unsupported endorsement types over 200 instances.

ment through topic continuation are harder to suppress because they lack the explicit agreement markers that the critic’s NLI-based scoring targets. Discourse-structure-aware scoring is the direct extension indicated by this residual category. Elaborative expansion remains the second-largest residual category, indicating that downstream amplification often happens through adding detail rather than direct repetition. The residual implicit-endorsement rate is 4.2%; topic continuation can preserve false presuppositions without overt agreement markers.

P.8 Pipeline Audit

For the pipeline audit, auditor profiles score 1,600 model outputs (400 per method, 100 per benchmark). The stratified audit oversamples positive and near-threshold outputs; parenthetical automatic rates in Table 27 are recomputed on this enriched subset and therefore differ from the full-test canonical rates in Table 1.

The audit gives precision ≥ 0.90 across all methods. The audit-adjusted rates are 1.6% for SOCIALCRITIC and 7.6% for LLM-only.

P.9 DealOrNoDeal Utility Analysis

Agreement alone can hide poor negotiation outcomes, so we also score expected utility on DealOrNoDeal. SOCIALCRITIC has expected utility 8.6 versus 7.4 for SAFE-Dialogue, 6.9 for verification prompting, and 6.2 for LLM-only, with endorsement rate 1.5%. Accepted-deal rates are 72.4% versus 70.8% for SAFE-Dialogue.

Metric	Prec.	Recall	F1	κ	Audit-adj. rate
Unsupported claim (SC)	0.94	0.89	0.91	0.82	5.4% (vs. 5.7% auto)
Unsup. endorsement (SC)	0.96	0.87	0.91	0.84	1.6% (vs. 1.7% auto)
Unsupported claim (SAFE-D)	0.92	0.87	0.89	0.80	8.2% (vs. 8.4% auto)
Unsup. endorsement (SAFE-D)	0.94	0.85	0.89	0.81	3.8% (vs. 4.0% auto)
Unsupported claim (Verif.)	0.90	0.84	0.87	0.77	12.4% (vs. 12.8% auto)
Unsup. endorsement (Verif.)	0.92	0.83	0.87	0.78	5.2% (vs. 5.4% auto)
Unsupported claim (LLM)	0.91	0.86	0.88	0.79	15.8% (vs. 16.1% auto)
Unsup. endorsement (LLM)	0.93	0.84	0.88	0.80	7.6% (vs. 7.9% auto)

Table 27: Audit using auditor profiles over 1,600 outputs.

P.10 Episode-Level Distribution

Figure 5 shows how episode-level actionable accuracy varies across the paired test distribution rather than only at the aggregate mean.

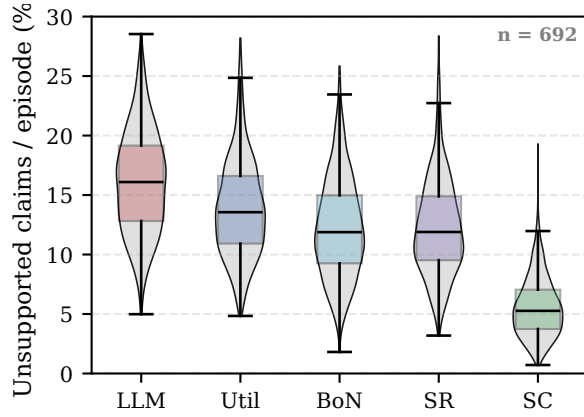


Figure 5: Episode-level distribution across four dialogue benchmarks (692 episodes).

The distribution view complements the benchmark averages in the main paper. It shows that the leftward shift is broad rather than concentrated in a handful of outlier episodes. The lower tail also becomes denser, which is consistent with more episodes staying near zero unsupported claims.

P.11 Critic Intrinsic Validation: Per-Benchmark Breakdown

Table 28 provides per-benchmark intrinsic quality for each critic.

The table shows that downstream benchmark difficulty and intrinsic critic quality move in the same direction. This alignment helps explain why SOTOPIA remains the most challenging benchmark even after critic guidance. The per-benchmark breakdown confirms that the aggregate critic scores reflect all four evaluation domains.

P.12 Backbone Generalization: Full Results

Table 29 reports all five metrics for the backbone-transfer experiment on LIGHT and SOTOPIA.

The full table shows that the ranking is preserved for all five reported metrics. The absolute degradation under the Llama planner is shared by every method, while the SOCIALCRITIC advantage remains stable across both evaluated domains.

Critic	Metric	LIGHT	DoND	SOTOPIA	P-C
Utility	Pairwise rank acc.	80.2%	76.8%	74.6%	81.8%
Norm	F1 at $\tau=-0.1$	0.91	0.87	0.86	0.92
Epistemic	F1 at $\tau=-0.1$	0.87	0.83	0.82	0.88

Table 28: Per-benchmark critic quality.

Backbone	Method	GoalComp $\uparrow$	AA $\uparrow$	UnsupClaim $\downarrow$	EndorseRate $\downarrow$	NormViol $\downarrow$
<i>LIGHT (150 episodes, 10 seeds)</i>						
GPT-4o-mini	LLM-only	62.7 ± 1.6	50.4 ± 2.1	17.4 ± 0.8	8.6 ± 0.6	17.2 ± 1.1
GPT-4o-mini	SAFE-Dialogue	75.8 ± 1.2	70.2 ± 1.4	8.6 ± 0.5	3.4 ± 0.3	7.6 ± 0.6
GPT-4o-mini	SOCIALCRITIC	82.6 ± 1.0	78.4 ± 1.1	6.4 ± 0.4	1.9 ± 0.2	3.8 ± 0.4
Llama-3.1-70B	LLM-only	59.4 ± 1.8	46.8 ± 2.3	18.8 ± 0.9	9.4 ± 0.7	18.8 ± 1.2
Llama-3.1-70B	SAFE-Dialogue	72.4 ± 1.3	66.8 ± 1.5	9.4 ± 0.5	3.8 ± 0.3	8.4 ± 0.7
Llama-3.1-70B	SOCIALCRITIC	79.8 ± 1.1	75.6 ± 1.2	6.8 ± 0.4	2.2 ± 0.2	4.2 ± 0.4
<i>SOTOPIA (90 episodes, 10 seeds)</i>						
GPT-4o-mini	LLM-only	54.8 ± 2.0	41.6 ± 2.4	19.6 ± 1.0	10.2 ± 0.7	22.4 ± 1.4
GPT-4o-mini	SAFE-Dialogue	68.4 ± 1.6	62.2 ± 1.8	9.8 ± 0.6	4.2 ± 0.4	10.2 ± 0.8
GPT-4o-mini	SOCIALCRITIC	75.4 ± 1.3	71.2 ± 1.5	7.4 ± 0.5	2.6 ± 0.3	4.6 ± 0.5
Llama-3.1-70B	LLM-only	51.2 ± 2.2	38.4 ± 2.6	21.4 ± 1.1	11.8 ± 0.8	24.6 ± 1.5
Llama-3.1-70B	SAFE-Dialogue	65.2 ± 1.7	59.6 ± 1.9	10.8 ± 0.6	4.8 ± 0.4	11.6 ± 0.9
Llama-3.1-70B	SOCIALCRITIC	72.6 ± 1.4	68.4 ± 1.6	7.8 ± 0.5	3.1 ± 0.3	5.4 ± 0.5

Table 29: Backbone-generalization results.