WHEN UNSUPPORTED CLAIMS BECOME WORKING STATE: MEASURING RATIFICATION IN SEQUENTIAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

The same unsupported claim can be challenged and disappear in one agent trace, yet be accepted as a premise and redirect retrieval or action in another. We call the latter transition *ratification* and represent it as a claim–consumer state change, with separate injection, endorsement, and operational-reuse events and event-specific risk sets. This representation supports matched policy tests and ENDORSEGUARD, a plan selector whose epistemic critic predicts whether a candidate will carry unsupported content across the state-admission boundary. On 200 held-out Persuasion-Conflict scenarios, assertion-time and adoption-time guards yield 8.2% and 4.6% conditional propagation, respectively, under identical planner, detector, feedback, proposal, and refinement settings. The ordering survives blinded human adjudication (a 3.6-point gap), a component-disjoint DeBERTa/BGE evaluator (3.5 points), and direct annotation of all 500 TopiOCQA test dialogues (6.1% versus 4.1%). Under matched planner-call budgets, ENDORSEGUARD reduces Persuasion-Conflict propagation from MARCH's 4.1% to 2.1% while improving success from 78.4% to 81.0%; on TopiOCQA it raises answer F1 from Utility-only's 55.6 to 60.4, with the largest propagation reduction at topic switches (9.8% to 3.7%). Treating state admission as an explicit reliability event therefore reveals consequential contamination that response-local factuality misses and supplies a concrete control target for sequential agents.

1 INTRODUCTION

Sequential agents continually turn prior messages, observations, and tool outputs into the context for their next decision. This creates a reliability boundary after generation: an unsupported claim can be ignored, challenged, or corrected, but it can also be accepted as a premise that redirects retrieval, planning, tool calls, or embodied action. Response-local factuality assigns the same production-time error to all of these futures.

Consider the sentence "The deadline was extended to Friday." A request for a source preserves the system's uncertainty; rescheduling the work accepts the sentence into working state. We call this claim–consumer transition *ratification*. It is the event that gives an unsupported emission decision-bearing status.

Failure attribution locates where an error originated, response verification scores evidential support when content is produced, and trajectory metrics record whether an error remains present. Ratification asks a different question: which later computation adopted or operationally reused the claim? Answering it requires claim identity across turns, event-specific opportunities for adoption and reuse, and exclusions for correction, negation, quotation, and verification questions.

We instantiate this measurement object with claim clusters and three lifecycle events: unsupported *injection*, later *endorsement*, and affirmative repetition or operational reuse. Matched guards place the same detector and refinement machinery either at assertion time or at adoption time. ENDORSEGUARD then uses the lifecycle signal as an epistemic candidate-selection objective, allowing the evaluation object to test a concrete control consequence under matched planner calls.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

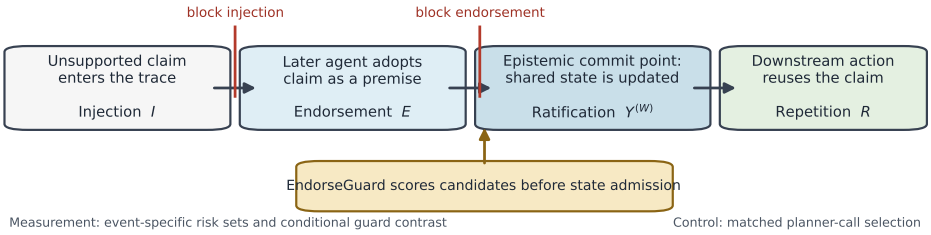

Figure 1: Ratification in a claim lifecycle. Injection I records unsupported entry; endorsement E and affirmative repetition or operational reuse R expose later state admission, producing $Y^{(W)}$. Paired guards compare injection-time and adoption-time veto policies while sharing the planner and refinement process. ENDORSEGUARD implements lifecycle-aware selection under the same planner-call budget.

Figure 1 connects the measured lifecycle to the two guard placements and lifecycle-aware control. The evidence follows the same sequence. On 200 held-out Persuasion-Conflict scenarios, assertion-time and adoption-time guards reach 8.2% and 4.6% conditional propagation under identical planner, detector, feedback, and refinement settings. The ordering persists under blinded adjudication, a component-disjoint evaluator, three temporal windows, and direct annotations of all 5,892 turns in the TopiOCQA test set, where the rates are 6.1% and 4.1%. As an end-to-end controller, ENDORSEGUARD reaches 2.1% propagation and 81.0% success, improving on MARCH at 4.1% and 78.4% with matched planner calls; on TopiOCQA, its largest reduction occurs at topic switches (9.8% to 3.7%).

Contributions. The primary contribution is a **state-transition measurement object** that formalizes injection, endorsement, repetition/reuse, event-specific risk sets, ratification windows, and semantic exclusions. The second is a **controlled mechanism study**: matched guard placements, independent human and model-based outcomes, temporal sensitivity, and complete public-dialogue annotations expose a stable conditional-propagation ordering. The third is a **commit-aware control result**: ENDORSEGUARD improves both contamination and task utility against training- and call-matched controllers, with corroborating gains in embodied collaboration and conversational retrieval. Together, these results identify state admission as a measurable reliability boundary and show that controlling it can improve downstream decisions without sacrificing task completion in the evaluated interfaces.

2 RELATED WORK

Ratification occupies a specific point in sequential computation: after a claim is emitted but before it governs a later response or action. Generation-time verification addresses the first boundary. Proposal selection, iterative refinement, process judges, and reinforced self-checking improve acceptance decisions for a candidate response (Cobbe et al., 2021; Lightman et al., 2024; Madaan et al., 2023; Dhuliawala et al., 2024; Li et al., 2026). TRACER links generated sentences to tool turns, evidence units, and semantic relations (Yu et al., 2026a), while ProvenanceGuard checks evidential support and source ownership in MCP traces (Alvarez et al., 2026). Ratification starts from that response-level support judgment and follows the claim to its next consumer: adoption, operational reuse, correction, and non-use are distinct outcomes even when the original emissions are equally unsupported.

Failure attribution and propagation methods observe other parts of the timeline. MIRAGE-Bench, Who&When, and AgentHallu identify a responsible agent or step (Zhang et al., 2025b;a; Liu et al., 2026); Causal Agent Replay estimates outcome-determining steps by counterfactual forward replay (Shah, 2026). Visual Flow measures evidence loss across turns (Yu et al., 2026b), HalluProp predicts pre-interaction risk from topology and role/query alignment (Lin et al., 2026), the Hallucination Snowball models error survival in linear pipelines (Singh & Pawar, 2026), and Schema combines concept-graph structure with trajectory severity (Feng et al., 2026). These methods locate origin, forecast risk, or summarize survival and consequence. Our claim–consumer unit resolves the intervening state change and assigns separate denominators to opportunities for injection, endorsement, and reuse.

Agent architectures make that state change operationally important. ReAct and Reflexion feed language-mediated reasoning or feedback into subsequent decisions (Yao et al., 2023; Shinn et al., 2023); Voyager and generative agents retain language in persistent skill or memory stores (Wang et al., 2023; Park et al., 2023). Such persistence mechanisms specify how prior content remains available, whereas ratification labels whether a particular unsupported claim is accepted as a premise by the next computation. The distinction separates mere presence in context from decision-bearing admission and applies to dialogue history, retrieved evidence, and embodied plans.

ENDORSEGUARD supplies a control realization of this event-factorized objective. Its utility, norm, and epistemic critics draw on calibration, retrieval grounding, attribution, RLHF, and Constitutional AI (Ji et al., 2023; Huang et al., 2025; Kadavath et al., 2022; Guo et al., 2017; Lewis et al., 2020; Guu et al., 2020; Rashkin et al., 2023; Gao et al., 2023; Ouyang et al., 2022; Bai et al., 2022). Unlike a binary task verifier, its epistemic target preserves unsupported emission and predicted downstream uptake as explicit labels; unlike MARCH, candidate selection receives a state-admission risk score under the common proposal/refinement interface. The evaluation draws on multi-agent debate, social simulation, and conversational IR (Du et al., 2024; Liang et al., 2024; Radlinski & Craswell, 2017; Dalton et al., 2020; Yu et al., 2021; Izacard & Grave, 2021; Adlakha et al., 2022; Anantha et al., 2021); the matched task-verifier and MARCH comparisons isolate the value of retaining the structured lifecycle signal at selection time.

3 RATIFICATION AS A TRACE-LEVEL STATE TRANSITION

3.1 COMMIT SEMANTICS AND CLAIM LIFECYCLE

An epistemic commit occurs when later computation admits an unsupported claim into working state by adopting it as a premise or reusing it in a subsequent decision. We expose this otherwise latent transition through observable claim-lifecycle events. Let trace $x_i = (o_1, a_1, \ldots, o_T, a_T)$ contain claim clusters $\mathcal{C}_i$, and let m_t denote the serialized dialogue, evidence, and action state visible before turn t. For claim c, $I_{ic} = 1$ when its first asserted mention lacks support in the evidence state available at that turn. For each eligible later turn t within window W, $E_{ict} = 1$ when the turn treats c as true while deriving a response or action, and $R_{ict} = 1$ when it carries c into shared context through an affirmative restatement or operational reuse. The labels may overlap when a turn both endorses and operationalizes a claim; ratification uses their union. Negation, correction, attributed quotation, a question requesting verification, and lexical repetition that neither affirms nor reuses the claim receive zero for both endpoints. Eligibility begins when the receiving agent can observe the injected mention and ends at the first correction, the trace boundary, or turn W. Ratification is

$$Y_{ic}^{(W)} = I_{ic} \mathbb{1}\left[\sum_{t=1}^{W} (E_{ict} + R_{ict}) > 0\right]. \tag{1}$$

Thus $Y_{ic}^{(W)}$ records whether claim c crosses the boundary within $W{=}5$ eligible turns (sensitivity: $W \in \{3, 8\}$). The audit reports injection over extracted clusters, propagation over unsupported clusters, and event rates over eligible later turns:

$$\widehat{\mathrm{Inj}} = \frac{\sum I}{\sum |\mathcal{C}|}, \quad \widehat{\mathrm{Prop}}_W = \frac{\sum Y^{(W)}}{\sum I}, \quad \widehat{\mathrm{EndEvt}} = \frac{\sum IE}{D_E}, \quad \widehat{\mathrm{RepEvt}} = \frac{\sum IR}{D_R}, \tag{2}$$

where D_E and D_R count turns eligible for the respective event. The cluster-conditioned and turn-conditioned denominators intentionally answer different questions. Two traces with equal response-local unsupportedness can therefore have different $\widehat{\mathrm{Prop}}_W$ when one claim becomes a premise for later action and the other is ignored.

The intervention study pairs every scenario and seed across two guard policies. *Block injection* vetoes a detected new unsupported assertion; *block endorsement* permits it but vetoes a detected later endorsement or premise-bearing reuse. Planner checkpoint, sampled scenario, random seed, detector, feedback template, $K{=}8$ proposals, and $R{=}2$ refinements are shared. For policy $z \in \{I, E\}$, the reported locus contrast is the difference between arm-specific conditional rates,

$$\widehat{\Delta}_{\mathrm{cond}} = \widehat{\mathrm{Prop}}_W^{(I)} - \widehat{\mathrm{Prop}}_W^{(E)}. \tag{3}$$

Each arm's denominator contains the unsupported clusters realized under that policy. The point estimate therefore describes a policy-conditional ordering rather than a claim-matched treatment effect. Scenario-paired bootstrap intervals resample $j = (\text{scenario}, \text{seed})$ and recompute both arm rates within each replicate, preserving their matched initial conditions while allowing their later trajectories and claim sets to differ.

3.2 Operational Detector and Guard Protocol

At every candidate turn, spaCy `en_core_web_lg` segments asserted sentences into claims. Agglomerative mention clustering uses `text-embedding-3-large`; a merge requires cosine distance at most 0.15 and bidirectional RoBERTa-large-MNLI entailment above 0.7. The same NLI model compares each new claim with every public or private evidence item visible to the speaking agent at that turn; a first asserted mention is unsupported when no visible item entails it above 0.7. For an unsupported cluster already visible in context, a later candidate is an endorsement when it entails the claim and uses it declaratively or as a premise. It is operational reuse when the claim supplies an argument to a plan, action, or query rewrite; an affirmative restatement is counted as repetition but a lexical mention without truth commitment is not. Contradiction or explicit correction closes the cluster's risk window. Negated mentions, attributed quotations, and evidence-seeking questions are excluded before event scoring.

A veto makes only the triggering proposal ineligible. Fixed-template feedback identifies the offending span and requests a supported reformulation or verification action, after which the same proposal slot is resampled; all other proposals remain unchanged. After at most two replacement rounds, the policy rescores the pool and selects the highest-scoring eligible candidate, using a fixed evidence-request action if no candidate passes the assigned guard. The injection arm applies this procedure only to a first unsupported assertion, whereas the endorsement arm applies it only when a later candidate affirms or operationally reuses an existing unsupported cluster. After the selected action, each environment continues normally from its own realized history; matched scenario and seed identify the statistical pair, not synchronized post-intervention claims.

3.3 Sequential Decision Process

We consider a partially observable social environment with latent state $s_t \in \mathcal{S}$ and actions $a_t \in \mathcal{A}$ spanning utterances, tool calls, and embodied actions. The environment transitions as $s_{t+1} \sim P(\cdot \mid s_t, a_t)$, while the agent-visible working state updates as $m_{t+1} = U(m_t, o_{t+1}, a_t)$. Ratification labels whether U preserves an unsupported cluster as an affirmed or operational premise. The process exposes task, norm, and epistemic rewards:

$$r_t^{\text{task}} \in \mathbb{R}, \qquad r_t^{\text{norm}} \in \{-1, 0\}, \qquad r_t^{\text{epi}} \in [-1, 0]. \tag{4}$$

We optimize

$$G_t = \sum_{k \geq 0} \gamma^k \left(r_{t+k}^{\text{task}} + \lambda_{\text{norm}} r_{t+k}^{\text{norm}} + \lambda_{\text{epi}} r_{t+k}^{\text{epi}} \right), \tag{5}$$

with $\gamma = 0.99$. The epistemic channel penalizes unsupported content in proportion to its predicted passage through the commit point into future working state.

3.4 EndorseGuard: Commit-Aware Control under Matched Budget

EndorseGuard equips planner π_θ with utility, norm, and epistemic critics. Each critic is a three-layer MLP head over a frozen Llama-3-8B-Instruct encoder (Llama Team, 2024) and consumes the same serialization of current observation, shared context, and delimited candidate plan. The utility head estimates goal progress; the norm head covers harm, deception, illicit guidance, privacy, and social-contract violations. The epistemic target is positive when a candidate introduces unsupported content, affirms or operationally reuses a prior unsupported cluster, or carries that cluster into a rollout decision within horizon H; correction, negation, attributed quotation, and evidence-seeking questions are negative. This composite target preserves support and ratification information instead of collapsing them into the trajectory-validity label used by the task-verifier control.

For candidate plan $\mathbf{a}_{t:t+H-1}$, each head emits $\hat{A}(s,a) \in [-1,1]$ and aggregates

$$A(\mathbf{a}_{t:t+H-1} \mid s_t) = \sum_{h=0}^{H-1} \gamma^h \hat{A}(\hat{s}_{t+h}, a_{t+h}), \tag{6}$$

where $\hat{s}_{t+h}$ comes from simulator rollouts or updates to textual and retrieved-evidence state. We use $H{=}4$. The shared score is

$$A^{\text{total}}(\mathbf{a} \mid s) = A_\phi^{\text{task}}(\mathbf{a} \mid s) + \lambda_{\text{norm}} A_\psi^{\text{norm}}(\mathbf{a} \mid s) + \lambda_{\text{epi}} A_\omega^{\text{epi}}(\mathbf{a} \mid s). \tag{7}$$

Task-family-specific heads preserve a common interface across collaboration, dialogue, and retrieval; Appendix A.12 evaluates cross-planner reuse within each task family.

Rollouts use only information available at the current prefix. Overcooked plans advance the standard environment simulator. Dialogue plans append candidate utterances and actions to the visible role, dialogue, and evidence serialization without synthesizing unobserved partner replies. Retrieval plans apply query rewrites, execute ConvDR against the frozen index, and write the returned top-10 passages into $\hat{s}$; ANSWERWITHCITATIONS and ASKCLARIFICATION terminate a candidate rollout. These task-family transition operators make the horizon score operational without reference answers or future test turns.

Algorithm 1 ENDORSEGUARD inference

Require: State s_t, planner π, critics $A_\phi^{\text{task}}, A_\psi^{\text{norm}}, A_\omega^{\text{epi}}$
Require: $K{=}8$, $R{=}2$, $\tau_{\text{norm}} = \tau_{\text{epi}} = -0.1$
Sample $\{\mathbf{a}^{(k)}\}_{k=1}^{K} \sim \pi(\cdot \mid s_t)$
for $r = 1$ to R **do**
  Score candidates and set $k^\star \leftarrow \arg\max_k A^{\text{total}}(\mathbf{a}^{(k)} \mid s_t)$
  if $A_\psi^{\text{norm}}(\mathbf{a}^{(k^\star)}) \geq \tau_{\text{norm}} \wedge A_\omega^{\text{epi}}(\mathbf{a}^{(k^\star)}) \geq \tau_{\text{epi}}$ **then**
    break
  end if
  $f \leftarrow \text{CriticHighlights}(\mathbf{a}^{(k^\star)})$
  Resample $\mathbf{a}^{(k^\star)} \sim \pi(\cdot \mid s_t, f)$
end for
Rescore all candidates and set $k^\star \leftarrow \arg\max_k A^{\text{total}}(\mathbf{a}^{(k)} \mid s_t)$
Return the first action of $\mathbf{a}^{(k^\star)}$

Nucleus sampling uses $p{=}0.9$ and $T{=}0.7$. `CriticHighlights` supplies fixed-template feedback for low-scoring spans, and the dual-threshold rule refines when either norm or epistemic risk remains. A_ϕ^{task} regresses normalized TD(λ) returns with $\lambda{=}0.95$. The norm and epistemic heads use class-balanced focal loss; a predicted risk probability p maps to advantage $1 - 2p$, so larger values consistently denote safer candidates in Eq. (7). This design places predicted state-admission risk directly in candidate selection while preserving a matched proposal/refinement interface.

ENDORSEGUARD uses at most ten planner generations and averages 8.7 ± 0.4 calls per decision step. These calls comprise one parallel eight-proposal round and 0.7 sequential refinements on average. The final rescore is local critic computation and adds no planner call. Appendix A.3 reports critic compute separately and decomposes the measured 731 ± 62 ms end-to-end latency on one A100 40GB GPU.

4 EXPERIMENTS

The experiments follow three questions. **RQ1** asks whether assertion-time and adoption-time guards produce different conditional propagation under matched initial conditions and machinery. **RQ2** tests whether the lifecycle ordering persists under independent evaluators, temporal windows, and a fully annotated public conversational workload. **RQ3** asks whether using state-admission risk in candidate selection improves the utility–contamination frontier against training- and call-matched controllers. Table 1, Appendix Table 6, and Table 2 answer RQ1–RQ2; Appendix Table 7, Overcooked-AI, and QReCC answer RQ3 under dialogue, embodied, and retrieval state interfaces.

4.1 EVALUATION SETTINGS

For Overcooked-AI, we use three standard layouts (Carroll et al., 2019): `Cramped Room`, `Asymmetric Advantages`, and `Coordination Ring`. We evaluate 1,500 episodes over 10 seeds; each episode runs for 400 timesteps, with a planner invocation every 5 timesteps.

Persuasion-Conflict comprises 800 controlled scenario specifications covering negotiation with hidden information, conflict de-escalation, collaborative fact-checking where sources contradict one another, and ethical persuasion subject to explicit constraints. Instead of providing a fixed transcript, each item defines the task goal and roles, private facts and public evidence, admissible interaction constraints, and a turn budget. The reported test split holds out 200 scenarios (50 per category); leakage control operates at the scenario-specification level. Critic fitting, validation, threshold calibration, and weak-supervision development use the other 600 scenarios, excluding every test role description, private fact, and evidence bundle from these stages. Scenario identifiers are partitioned before trace generation, and every seeded realization of an identifier remains in the same split. Examples include accepting an unsupported reservation price, treating an unverified de-escalation claim as common ground, carrying a contradicted source into fact-checking, and reusing an impermissible persuasive premise. At each turn, the item's public fields plus the speaking agent's permitted private fields define the evidence state for support labels. Success requires completing the declared goal within the turn budget; an invalid transition violates an item-level role or interaction constraint, and norm violations follow the five-category taxonomy in Appendix A.2. This contract makes injection, endorsement, repetition, correction, and negation directly labelable while paired scenario–seed runs preserve the same initial task conditions.

TopiOCQA (Adlakha et al., 2022) has 500 test dialogues (5,892 turns) with topic switching over a 25.7M-passage Wikipedia corpus. We add direct ratification annotations to every test dialogue at 3-, 5-, and 8-turn windows. 3 blinded labels per claim cluster distinguish unsupported assertion, adoption as a premise, repetition/reuse, correction, negation, attributed quotation, and question rewrite. Majority adjudication produces the primary endpoints; topic-switch and same-topic turns are retained as prespecified slices. QReCC (Anantha et al., 2021) contains 2,775 test turns over a 54M-passage Natural Questions corpus. TREC CAsT 2020 (Dalton et al., 2020) contains 25 topics (216 turns) over an MS MARCO + CAR corpus (38.8M passages). All methods use ConvDR (Yu et al., 2021) as the base retriever. The planner chooses among REWRITEQUERY, RETRIEVE, ANSWERWITHCITATIONS, and ASKCLARIFICATION. The IR agent loop starts with a short action-plan proposal, followed by optional ConvDR retrieval of top-k=10 passages and candidate scoring against the resulting evidence set. Retrieval quality and answer F1 receive rewards from the utility critic, whereas evidence support, citation alignment, and predicted downstream adoption are scored by the epistemic critic.

4.2 INDEPENDENT OUTCOME EVALUATION

We freeze outputs from the LLM-only reference, both veto interventions, Utility-only, and ENDORSE-GUARD on all 200 Persuasion-Conflict test scenarios and 10 seeds. A category- and method-stratified sample of 1,260 claim clusters receives 3 independent labels per cluster. Annotators see the dialogue, private/public evidence available at each turn, and a randomized claim identifier; method identity, original detector output, and intervention condition are concealed. Majority labels define support, endorsement, and repetition, with agreement and disagreement reported in Table 6. Human rates use inverse method–category inclusion weights; scenario-paired bootstrap replicates retain every selected cluster and majority label linked to a sampled scenario. Appendix A.4 gives the estimator details and clarifies that policy-specific cluster frames are not post-hoc claim matched.

The second outcome evaluator shares no component with the veto detector. Stanza `en_ewt` supplies sentence and dependency spans, `BAAI/bge-large-en-v1.5` clusters claims at cosine similarity 0.82, and `MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli` assigns support and entailment at probability 0.74. Its thresholds are fixed on non-test validation dialogues, and its outputs never trigger proposal rejection. This design isolates the reported outcome from the intervention mechanism while preserving the same frozen traces and paired scenario/seed units.

4.3 BASELINES

We compare against LLM-only, Best-of-N with a GPT-4-turbo judge, verification prompting, Self-Refine (Madaan et al., 2023), Utility-only, and rule-based filtering. Appendix A.3 reports exact call counts. Two close learned controls use the identical held-out split, GPT-4o-mini planner, proposal/refinement interface, tuning access, and training-label budget. *Task verifier* places a single accept/reject head on the same frozen Llama-3-8B encoder and collapses support, endorsement, and norm labels into one trajectory-validity objective. *MARCH* implements reinforced multi-agent self-checking (Li et al., 2026): its Solver emits common-language plans, its Proposer and Checker return claim-level verification through the shared refinement channel, and the accepted plan contributes its first action. Both receive up to ten planner generations; verifier scoring and self-check compute are timed separately. This comparison isolates event-factorized ratification-aware selection from a benchmark-tuned binary verifier and a recent reinforced self-check controller.

Additional conversational-retrieval baselines are Best-of-N (IR), Utility-only (IR), ConvDR+FiD, and an LLM reranker. Controller comparisons use the same five-call IR budget, corpus, index, retrieval depth, and answer generator.

4.4 METRICS

For task performance, Success Rate records goal completion irrespective of constraint validity, while Score covers Overcooked-AI and Recall@10/MRR@10/answer F1/nDCG@3 cover conversational IR. Actionable Accuracy (AA) counts the proportion of episodes in which the task goal is achieved and neither invalid transitions nor norm violations occur:

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

NormViol is the fraction of episodes with a norm violation; Unsup./Inj. is the unsupported extracted-cluster rate. Propagation, EndEvt., and RepEvt. follow Eq. (2). The guard-locus contrast in Eq. (3) is positive when endorsement blocking has the lower arm-specific conditional propagation rate. Controller and TopiOCQA complete-trace percentages have raw numerators and denominators in the main table or Appendix A.6; the independently labeled human sensitivity reports inverse-stratum-weighted rates from its declared 1,260-cluster frame. Means and standard errors are computed across seeds. Confidence intervals use 10,000 replicates of paired scenarios or dialogues within seed; no turn or claim is treated as an independent unit.

The original Persuasion-Conflict evaluator also triggers the targeted vetoes. For sentence-by-sentence claim extraction, we use the large English pipeline from spaCy (`en_core_web_lg`). To merge co-referent spans, agglomerative embedding clustering uses `text-embedding-3-large` with cosine distance at most 0.15, together with RoBERTa-large-MNLI as an NLI gate requiring bidirectional entailment > 0.7. Within a 5-turn window, entailment of an earlier unsupported claim by a later turn is detected as endorsement. We calibrate weak supervision with 2,400 human-labeled norm samples ($\kappa = 0.83$) and 1,800 human-labeled claim-support samples ($\kappa = 0.79$), and Appendix A.4 reports the full 400-claim manual audit. The independent human and DeBERTa/BGE evaluations directly test evaluator sensitivity; Appendix A.9 crosses evaluators, windows, and clustering/NLI thresholds.

4.5 IMPLEMENTATION DETAILS

GPT-4o-mini (`gpt-4o-mini-2024-07-18`, temperature 0.7, top-p 0.9), accessed through the OpenAI API, supplies all planner proposals and self-improvement baselines unless otherwise noted; `gpt-4-turbo-2024-04-09` supplies the judge in the Best-of-N baseline. For the critics, frozen `meta-llama/Meta-Llama-3-8B-Instruct` encoders feed 3-layer MLP heads whose hidden widths are 1024, 512, 256. The default settings assign $\lambda_{\text{norm}} = \lambda_{\text{epi}} = 0.5$ to the weights, $K = 8$ to the proposal count, and $H = 4$ to the horizon, allowing at most $R = 2$ refinements; a 5-call budget in conversational IR is maintained by using K=4 and R=1. Mixed-policy training contains 36,000 Overcooked trajectory windows, 18,000 Persuasion-Conflict traces, and 27,500 conversational-IR traces; disjoint validation sets contain 4,500, 2,400, and 3,200, respectively. IR traces come from dialogue-disjoint official TopiOCQA and QReCC training splits; no test dialogue, annotation, or

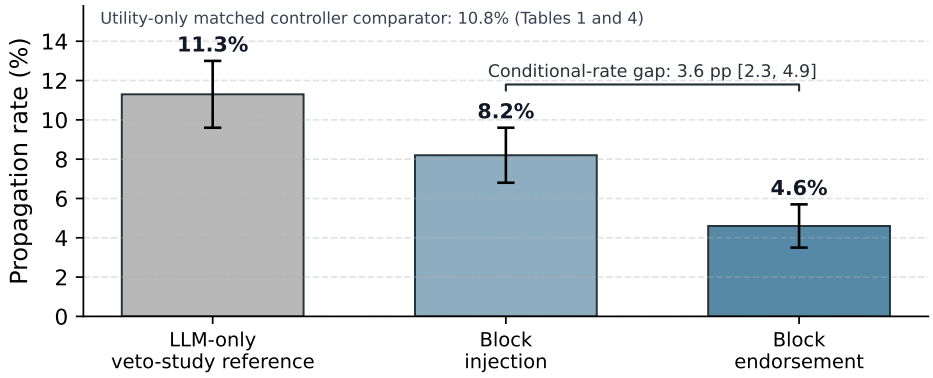

Figure 2: Persuasion-Conflict guard policies. Rates condition on each policy's realized unsupported clusters; error bars and the 3.6-point gap use scenario-paired bootstrap 95% CIs.

CAsT topic enters fitting. Heads use AdamW at 2e-4, batch size 128, class-balanced focal loss, and validation AUPRC early stopping; thresholds maximize validation AA subject to non-increasing propagation. All learned controls receive the same traces, label budget, validation access, and planner checkpoints. Training uses $4\times$ A100 80GB GPUs and inference one A100 40GB GPU; seed counts, loss details, and calibration appear in Appendix A.8.

5 Results

5.1 Adoption-Time Blocking Produces Lower Conditional Propagation

The paired veto study compares two concrete placements of the same detector and refinement machinery. Figure 2 uses the 11.3% LLM-only veto-study reference, while the matched controller comparison in Tables 1 and 7 uses Utility-only at 10.8%. With planner, feedback, detector, and $K{=}8$, $R{=}2$ refinement held fixed, injection blocking reaches 8.2% propagation (-27% relative), whereas endorsement blocking reaches 4.6% (-59% relative). The arm-specific conditional-rate gap is 3.6 points; a scenario-paired bootstrap gives a 95% CI of [2.3,4.9].

Table 1 shows how event-factorized control changes the full claim lifecycle. Relative to Utility-only, EndorseGuard reduces unsupported injection by 7.7 points (12.9% to 5.2%) and propagation by 8.7 points (10.8% to 2.1%). Endorsement events fall from 6.4% to 1.4%, repetition falls from 4.4% to 0.7%, and AA rises by 21.2 points in Table 7. The joint movement of endorsement, repetition, and AA shows that lower propagation accompanies less downstream uptake and stronger constraint-respecting task completion.

Table 1: Persuasion-Conflict claim lifecycle, mean (s.e.) over 10 seeds. Prop. conditions on unsupported clusters; Inj. on extracted clusters; event rates on eligible later turns.

Method	Prop. ↓	Inj. ↓	EndEvt. ↓	RepEvt. ↓
LLM-only	11.3 (0.9)	14.2 (1.1)	6.8 (0.7)	4.5 (0.5)
Best-of-N (matched)	8.7 (0.8)	11.4 (0.8)	5.2 (0.6)	3.5 (0.4)
Verif. prompt (matched)	6.2 (0.6)	8.6 (0.7)	3.4 (0.4)	2.1 (0.3)
Utility-only	10.8 (0.9)	12.9 (1.0)	6.4 (0.7)	4.4 (0.5)
EndorseGuard	2.1 (0.3)	5.2 (0.5)	1.4 (0.2)	0.7 (0.1)

5.2 The Guard Ordering Survives Evaluator Replacement

Appendix Table 6 evaluates frozen traces with outcome evaluators isolated from veto decisions. Human adjudication measures 10.7% propagation for the frozen reference, 7.9% under injection

blocking, and 4.3% under endorsement blocking. The disjoint DeBERTa/BGE outcome stack shifts absolute rates to 11.8%, 8.5%, and 5.0%, while preserving a 3.5-point rate gap under scenario-paired bootstrap. The same evaluators give EndorseGuard–Utility-only margins of 7.6 and 8.3 points (Appendix A.4).

Human agreement is 0.81 for support, 0.78 for endorsement, and 0.75 for repetition (Krippendorff's α), with 13.4% three-way disagreement before majority adjudication. The four scenario categories contribute 318, 306, 327, and 309 clusters. The preserved ordering across independent semantics, models, and annotators shows that the result is not tied to the original detector's scoring semantics.

5.3 Public Conversational Traces Reproduce the Guard Ordering

The complete TopiOCQA test set retains the intervention ordering under direct claim-lifecycle annotation. Across 5,892 turns, EndorseGuard reaches 3.4% propagation, versus 7.5% for the matched Utility-only controller, while increasing answer F1 from 55.6 to 60.4. The intervention comparison again favors endorsement blocking: 4.1% versus 6.1%, a 2.0-point conditional-rate gap (scenario-paired 95% CI [0.9,3.1]). Question rewrites count as repetition when they preserve the claim and as endorsement when they use it as a premise for a new answer or retrieval action. As in Persuasion-Conflict, the displayed denominators contain the unsupported clusters realized under each policy; the paired interval resamples dialogues and preserves their common initial questions.

Table 2: Direct ratification on all 500 TopiOCQA test dialogues. Prop. gives ratified/unsupported counts; means are over five seeds.

Method	Prop. % (n/N) ↓	End. ↓	Rep. ↓	F1 ↑	R@10 ↑	Calls
LLM-only (shared retrieval)	8.6 (93/1,081)	4.5	3.1	51.2	61.4	1
LLM rerank.	7.1 (75/1,057)	3.8	2.5	56.8	67.2	4
Best-of-N (IR)	6.9 (74/1,072)	3.6	2.4	55.8	66.2	5
Utility-only (IR)	7.5 (78/1,040)	4.0	2.6	55.6	65.8	5
EndorseGuard	3.4 (34/1,003)	1.8	1.1	60.4	70.6	5
Block injection	6.1 (64/1,049)	3.3	2.0	58.3	68.5	5
Block endorsement	4.1 (42/1,024)	2.2	1.3	59.1	69.3	5

Topic-switch turns show the largest matched-controller gain: EndorseGuard reduces propagation from 9.8% to 3.7%; same-topic turns move from 6.4% to 3.1%. This heterogeneity links ratification risk to context transitions, where an adopted premise is most likely to distort the next retrieval state.

5.4 Commit-Aware Control Improves the Utility–Contamination Frontier

Appendix Table 7 compares end-to-end controllers. The task-specific verifier and MARCH provide strong controls at 4.9% and 4.1% propagation. EndorseGuard further reaches 2.1%, with paired margins of 2.8 points (CI [1.6,4.0]) and 2.0 points (CI [0.9,3.1]). It also improves success by 4.2 points over the task verifier and 2.6 over MARCH, so the propagation margin is achieved on a stronger utility frontier.

6 Conclusion

Ratification isolates the transition by which unsupported content becomes a premise for later computation. Claim-level injection, endorsement, and operational-reuse labels make this state change measurable across dialogue and retrieval traces: adoption-time blocking reaches 4.6% conditional propagation versus 8.2% for assertion-time blocking on Persuasion-Conflict, with the ordering preserved by independent evaluators and direct TopiOCQA annotations. Under matched planner calls, EndorseGuard reaches 2.1% propagation and 81.0% success while improving retrieval and collaboration outcomes. Sequential-agent evaluation should therefore track not only what a model asserts, but which assertions acquire authority over subsequent state updates; that boundary supplies both a sharper diagnosis of shared-context contamination and a concrete point for control.

AI USE STATEMENT

Generative AI tools were used solely to edit and improve the manuscript's writing and presentation, including the system name, title, abstract, and introduction. They were not used to formulate the research ideas, methods, experiments, data, mathematical claims, proofs, code, analyses, or interpretations. All AI-assisted edits were reviewed by the authors, who take responsibility for the final content.

ETHICS STATEMENT

Persuasion-Conflict uses controlled role and scenario specifications, and annotators evaluate agent dialogues without interacting with deployed systems or supplying private conversations. The norm taxonomy targets harm, deception, illicit guidance, privacy violations, and social-contract violations; annotators may nevertheless differ across cultural settings, so deployment-specific calibration remains important. The study's threat model includes manipulative and unsupported language, which could inform misuse as well as defenses. Access control, data minimization, red-team evaluation, and human oversight are appropriate safeguards for high-stakes applications of ENDORSEGUARD.

REPRODUCIBILITY STATEMENT

Sections 3–5 specify the estimands, models, training boundaries, evaluation splits, paired uncertainty procedure, and planner-call accounting. Appendices A.1–A.12 provide utility transfer, latency, annotation, sensitivity, calibration, raw-count, stress-test, and cross-planner details.

REFERENCES

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483, 2022. doi: 10.1162/tacl_a_00471. URL https://aclanthology.org/2022.tacl-1.27/.

Ander Alvarez, Santhiya Rajan, Alessandro Genuardi, Oliver Wirjadi, Samuel Mugel, and Román Orús. ProvenanceGuard: Source-aware factuality verification for MCP-based LLM agents, 2026. URL https://arxiv.org/abs/2606.18037.

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 520–534. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.44. URL https://aclanthology.org/2021.naacl-main.44/.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022. doi: 10.48550/arXiv.2212.08073. URL https://arxiv.org/abs/2212.08073.

Micah Carroll, Rohin Shah, Mark K. Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-AI coordination. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 5175–5186, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f5b1b89d98b7286673128a5fb112cb9a-Abstract.html.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168. URL https://arxiv.org/abs/2110.14168.

Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. CAsT 2020: The conversational assistance track overview. In *Proceedings of the Twenty-Ninth Text REtrieval Conference (TREC)*. NIST, 2020. URL https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.C.pdf.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.212. URL https://aclanthology.org/2024.findings-acl.212/.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 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*, pp. 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Xinshun Feng, Ziqi Miao, Lijun Li, and Jing Shao. Tracing the cascade: A topology-aware evaluation framework for scientific agent hallucinations, 2026. URL https://arxiv.org/abs/2608.00711.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 2020. URL https://proceedings.mlr.press/v119/guu20a.html.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 874–880. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74/.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. doi: 10.48550/arXiv.2207.05221. URL https://arxiv.org/abs/2207.05221.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Zhuo Li, Yupeng Zhang, Pengyu Cheng, Jiajun Song, Mengyu Zhou, Hao Li, Shujie Hu, Yu Qin, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. MARCH: Multi-agent reinforced check for hallucination. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 39389–39415. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1828. URL https://aclanthology.org/2026.acl-long.1828/.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2305.20050.

Shi Lin, Chenpei Wang, Peng Qian, Dezhang Kong, Minghao Li, Yufeng Li, and Xun Wang. Before agents speak: Pre-hoc failure risk inference in multi-agent systems, 2026. URL https://arxiv.org/abs/2607.26836.

Xuannan Liu, Xiao Yang, Zekun Li, Peipei Li, and Ran He. AgentHallu: Benchmarking automated hallucination attribution of LLM-based agents. *arXiv preprint arXiv:2601.06818*, 2026. doi: 10.48550/arXiv.2601.06818. URL https://arxiv.org/abs/2601.06818.

Llama Team. The Llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783.

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. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2303.17651.

OpenAI. GPT-4 technical report. Technical report, OpenAI, 2023. URL https://arxiv.org/abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:27730–27744, 2022. URL https://arxiv.org/abs/2203.02155.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM, 2023.

Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Human Information Interaction and Retrieval (CHIIR)*, pp. 117–126. ACM, 2017.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023. doi: 10.1162/coli_a_00486. URL https://aclanthology.org/2023.cl-4.2/.

Jaineet Shah. Causal agent replay: Counterfactual attribution for LLM-agent failures, 2026. URL https://arxiv.org/abs/2606.08275.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://arxiv.org/abs/2303.11366.

Prabhjot Singh and Bhushan Pawar. The hallucination snowball: Modeling error propagation as state transitions in multi-agent LLM pipelines, 2026. URL https://arxiv.org/abs/2608.14588. Accepted at the ICML 2026 FAGEN Workshop.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Bihui Yu, Caijun Jia, Jing Chi, Xiaohan Liu, Yining Wang, He Bai, Yuchen Liu, Jingxuan Wei, and Junnan Zhu. TRACER: Verifiable generative provenance for multimodal tool-using agents, 2026a. URL https://arxiv.org/abs/2605.09934.

Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. Few-shot conversational dense retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 829–838. ACM, 2021. doi: 10.1145/3404835.3462856. URL https://doi.org/10.1145/3404835.3462856.

Xinlei Yu, Chengming Xu, Guibin Zhang, Yongbo He, Zhangquan Chen, Zhucun Xue, Jiangning Zhang, Yue Liao, Xiaobin Hu, Yu-Gang Jiang, and Shuicheng Yan. Visual multi-agent system: Mitigating hallucination snowballing via visual flow. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=Ii4HBlERix.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, and Qingyun Wu. Which agent causes task failures and when? on automated failure attribution of LLM multi-agent systems. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76583–76599. PMLR, 2025a. URL https://proceedings.mlr.press/v267/zhang25cq.html.

Weichen Zhang, Yiyou Sun, Pohao Huang, Jiayue Pu, Heyue Lin, and Dawn Song. MIRAGE-Bench: LLM agent is hallucinating and where to find them. *arXiv preprint arXiv:2507.21017*, 2025b. doi: 10.48550/arXiv.2507.21017. URL https://arxiv.org/abs/2507.21017.

A.1 OVERCOOKED-AI SUPPORTING USEFULNESS TABLE

Table 3 reports the collaboration results summarized in the main text. This setting tests whether commit-aware selection preserves useful joint action under a different state and action interface; Persuasion-Conflict and TopiOCQA provide the direct claim-lifecycle endpoints.

Table 3: Overcooked-AI results (1,500 episodes, three layouts). Utility-only is the matched planner-call baseline. Entries are mean (s.e.) over 10 seeds.

Method	Succ. ↑	AA ↑	Score ↑	NormViol ↓	Calls
LLM-only	72.4 (2.1)	58.6 (2.5)	3.42 (0.18)	18.2 (1.4)	1.0
Self-Refine	78.1 (1.8)	67.2 (2.2)	3.81 (0.15)	11.4 (1.1)	4.0
Best-of-N (matched)	79.6 (1.7)	68.4 (2.1)	3.92 (0.14)	12.6 (1.0)	9.0
Verif. prompt (matched)	81.2 (1.6)	71.8 (1.9)	4.08 (0.13)	9.8 (0.9)	9.0
Rule-based	64.8 (2.4)	60.2 (2.3)	2.94 (0.21)	5.4 (0.7)	11.2
Utility-only	76.4 (1.9)	65.1 (2.2)	3.68 (0.16)	14.2 (1.2)	8.9
ENDORSEGUARD	89.7 (1.2)	84.6 (1.4)	4.82 (0.11)	4.1 (0.5)	8.7

A.2 NORM TAXONOMY

We categorize norm violations as harm, deception, illicit guidance, privacy, or social-contract violations following established safety taxonomies (OpenAI, 2023; Bai et al., 2022). Each label also records minor, moderate, or severe impact.

A.3 PLANNER-CALL ACCOUNTING AND LATENCY

All matched controllers begin with the same proposal budget. ENDORSEGUARD averages fewer refinements than Utility-only, and critic scoring is reported as local auxiliary compute. The 8.7 average planner calls comprise one parallel proposal round and 0.7 sequential refinements. The final all-candidate rescore in Algorithm 1 is included in critic latency and does not invoke the planner.

Table 4: Planner LLM calls per decision step. Entries are mean (s.e.) where applicable.

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
Verif. prompting (matched)	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)
ENDORSEGUARD	8	0	0.7 (0.4)	8.7 (0.4)

The measured wall time is $t_{\text{step}} \approx t_{\text{proposal-round}} + 0.7t_{\text{gen}} + t_{\text{critic}} + t_{\text{env}}$. The initial parallel proposal round plus dispatch and aggregation accounts for the difference between the component means and the 731 ms total.

A.4 EPISTEMIC ANNOTATION AUDITS

The original evaluator extracts claims with spaCy's large English pipeline. It clusters representations from `text-embedding-3-large` at cosine threshold 0.15 and applies bidirectional RoBERTa-large-MNLI entailment above 0.7. A 400-claim audit yields 3.8% segmentation error, 2.5% support error, and 4.8% endorsement-detection error. Because stage errors overlap, 7.2% of the same 400 claims contain at least one pipeline-stage error; this is the reported aggregate end-to-end rate

Table 5: Latency on one A100 40GB GPU with parallel proposal calls.

Component	Latency (ms)	Std.
GPT-4o-mini API (one generation)	312	48
GPT-4-turbo API (one judge call)	687	112
Critic scoring (eight candidates)	96	8
Simulator step (Overcooked-AI)	2.1	0.3
ENDORSEGUARD total (8.7 calls/step)	731	62

rather than a sum of stage rates. Assigning every audited error against ENDORSEGUARD raises its Persuasion-Conflict propagation from 2.1% to 3.6%, a 68% reduction relative to LLM-only.

Table 6: Independent evaluation of frozen veto traces. Columns are arm-specific conditional rates; differences use scenario-paired bootstrap 95% CIs. Neither evaluator participates in veto decisions.

Outcome evaluator	LLM ref.	Block Inj.	Block End.	Inj.−End.
Human majority	10.7 (0.8)	7.9 (0.7)	4.3 (0.5)	3.6 [2.2,5.0]
DeBERTa/BGE	11.8 (0.9)	8.5 (0.7)	5.0 (0.5)	3.5 [1.9,5.1]

Table 7: Persuasion-Conflict controllers on 200 held-out scenarios, mean (s.e.) over 10 seeds. Proposal/refinement budgets are matched.

Method	Succ. ↑	AA ↑	NormViol ↓	Unsup. ↓	Prop. ↓	Calls
LLM-only	62.5 (3.1)	47.8 (3.6)	21.5 (2.1)	14.2 (1.1)	11.3 (0.9)	1.0
Self-Refine	69.5 (2.7)	58.2 (3.1)	14.5 (1.7)	10.8 (0.9)	8.4 (0.8)	4.0
Best-of-N (matched)	68.5 (2.8)	56.8 (3.2)	15.8 (1.8)	11.4 (0.8)	8.7 (0.8)	9.0
Verif. prompt (matched)	70.5 (2.6)	60.4 (2.9)	13.2 (1.5)	8.6 (0.7)	6.2 (0.6)	9.0
Rule-based	55.0 (3.4)	51.6 (3.3)	6.5 (1.0)	13.8 (1.0)	10.6 (0.9)	11.2
Utility-only	67.5 (2.9)	55.6 (3.1)	16.8 (1.9)	12.9 (1.0)	10.8 (0.9)	8.9
Task verifier	76.8 (2.2)	70.9 (2.5)	8.1 (1.1)	6.9 (0.6)	4.9 (0.5)	8.8
MARCH	78.4 (2.1)	73.1 (2.3)	6.7 (0.9)	6.4 (0.6)	4.1 (0.4)	8.9
ENDORSEGUARD	81.0 (1.9)	76.8 (2.1)	4.5 (0.8)	5.2 (0.5)	2.1 (0.3)	8.7

Appendix Table 17 isolates complementary objectives: adding the epistemic critic reduces propagation from 10.8% to 4.2%, versus 8.4% with the norm critic, while their combination reaches 2.1% propagation and 81.0% success.

Beyond Persuasion-Conflict, ENDORSEGUARD raises Overcooked success from 76.4% to 89.7% and QReCC F1 from 52.6 to 57.2 over Utility-only while reducing norm violations or unsupported claims (Appendices A.1 and A.12).

Interpretation and scope. Across evaluator replacement, three temporal windows, six held-out stress categories, three planners, and complete public conversational traces, lower downstream uptake consistently accompanies stronger task outcomes. The guard-locus rates answer a policy-conditional question: each intervention changes its later trajectory and unsupported-claim risk set, while paired bootstrap preserves common starting scenarios rather than constructing claim-level counterfactuals after divergence. The adaptive suite localizes the main residual errors to delayed reuse and correction/negation, and the public result is strongest at topic switches, where an admitted premise immediately shapes the next retrieval state. Within the evaluated dialogue, retrieval, and embodied interfaces, ratification is therefore an observable and controllable state transition rather than merely continued lexical presence.

For the independent study, 3 blinded annotators label each of 1,260 stratified claim clusters. The interface exposes the claim, prior dialogue, and turn-available evidence in randomized order. Annotators identify support, adoption as a premise, later repetition/reuse, correction, negation, and attributed

quotation. Each cluster receives 3 judgments, majority adjudication, and senior tie-breaking for 4.8% of clusters. The category totals are 318 negotiation, 306 de-escalation, 327 fact-checking, and 309 constrained-persuasion clusters. Within each method–category stratum, inverse inclusion weights map the labeled frame back to its frozen-trace cluster count. Bootstrap replicates resample scenario identifiers and retain all judgments linked to each selected scenario; because policies produce distinct cluster frames, no post-policy claim matching is assumed. On the matched-controller frames, human propagation is 10.2 (0.8) for Utility-only and 2.6 (0.4) for ENDORSEGUARD, a 7.6 [5.7,9.5]-point paired margin. The DeBERTa/BGE rates are 11.1 (0.9) and 2.8 (0.4), a 8.3 [6.3,10.3]-point margin. Compensation is $18/hour, and workers can withdraw before submission of each task. The alternative Stanza/BGE/DeBERTa evaluator is calibrated on development dialogues and remains isolated from intervention decisions.

A.5 BENCHMARK AND SPLIT DETAILS

Persuasion-Conflict uses 800 scenario specifications split at the scenario level: 200 held-out test scenarios and 600 development scenarios. Each specification defines a task goal, roles, private facts, public evidence, admissible interaction constraints, and a turn budget across negotiation, de-escalation, fact-checking, and constrained persuasion. The scenario identifier is the atomic partitioning unit: every seeded dialogue generated from one specification remains in one split. The 200 test specifications contain 50 items from each category, and held-out goals, role descriptions, private facts, and evidence bundles are excluded from critic fitting, weak-supervision development, and threshold calibration. For support scoring at turn t, the evidence state contains the specification's public evidence and only the private facts available to the current role. Success requires satisfying the declared task goal within the turn budget. An invalid transition violates an explicit role or interaction constraint, while norm violations use the harm, deception, illicit-guidance, privacy, and social-contract taxonomy in Appendix A.2. These fields provide the ground truth needed to distinguish unsupported assertions from contradictions in the available evidence and to adjudicate whether a later response treats a claim as established.

TopiOCQA contributes 500 test dialogues and 5,892 turns over its 25.7M-passage Wikipedia corpus. Direct ratification labels are attached to frozen outputs for all seven methods in Table 2. Each claim cluster receives 3 blinded judgments; Krippendorff's α is 0.80 for support, 0.76 for endorsement, and 0.73 for repetition. The adjudicated pool contains 311 endorsement events and 208 repetition events, with question rewrites separated from premise-bearing adoption. QReCC contributes 2,775 test turns over a 54M-passage Natural Questions corpus, and TREC CAsT 2020 contributes 25 topics and 216 turns over MS MARCO+CAR. All retrieval runs use ConvDR with a FAISS HNSW index.

A.6 RAW DENOMINATORS AND PAIRED UNCERTAINTY

Table 8 reports pooled claim and eligible-turn counts corresponding to the seed-level means in Table 1. Small rounding differences arise because the main table averages rates across seeds, whereas the appendix shows pooled raw counts. For the guard-policy comparison, Eq. (3) uses the unsupported clusters realized in each arm. Bootstrap replicates retain both outcomes for every resampled scenario–seed unit and recompute the two arm-level rates after resampling; no post-intervention claim matching is imposed.

Table 8: Persuasion-Conflict raw counts. Each cell is numerator/denominator for the metric in its column.

Method	Prop.	Inj.	EndEvt.	RepEvt.
LLM-only	100/884	884/6,225	189/2,779	125/2,779
Best-of-N	61/706	706/6,193	142/2,731	96/2,731
Verif. prompt	33/529	529/6,151	92/2,706	57/2,706
Utility-only	87/804	804/6,233	177/2,766	122/2,766
ENDORSEGUARD	7/319	319/6,135	37/2,643	19/2,643

Table 9: Representative paired bootstrap differences. Positive Success/AA and negative Propagation favor ENDORSEGUARD.

Comparison	Metric	Difference (95% CI)
EG vs. Best-of-N (Overcooked)	Success	+10.1 pp [+6.8,+13.4]
EG vs. Best-of-N (Overcooked)	AA	+16.2 pp [+12.6,+19.8]
EG vs. Verif. prompt (Overcooked)	AA	+12.8 pp [+9.6,+16.0]
EG vs. Utility-only (Persuasion)	Propagation	-8.7 pp [-10.4,-7.0]
EG vs. LLM-only (TopiOCQA)	R@10	+9.2 pp [+7.4,+11.0]
EG vs. task verifier (Persuasion)	AA	+5.9 pp [+2.4,+9.4]
EG vs. MARCH (Persuasion)	AA	+3.7 pp [+0.8,+6.6]

A.7 LEARNED-COMPARATOR DETAILS

The task verifier uses the frozen Llama-3-8B-Instruct encoder and the same training trajectories as ENDORSEGUARD, with a single binary accept/reject head. MARCH follows its reinforced multi-agent self-check protocol (Li et al., 2026); messages are mapped to the common action interface and planner generations are capped at ten. Threshold selection, planner checkpoint, proposal seeds, and held-out scenarios are shared.

Table 10: Training-matched Persuasion-Conflict comparison. Latency includes auxiliary verification or critic compute; planner calls remain separately matched.

Method	Succ.	AA	NormViol	Unsup.	Prop.	EndEvt.	Calls	ms
Task verifier	76.8	70.9	8.1	6.9	4.9	2.8	8.8	704 (55)
MARCH	78.4	73.1	6.7	6.4	4.1	2.3	8.9	952 (74)
ENDORSEGUARD	81.0	76.8	4.5	5.2	2.1	1.4	8.7	731 (62)

A.8 CRITIC TRAINING AND CALIBRATION

Training uses mixed-policy successful and near-miss traces with class-balanced norm and epistemic minibatches. Task returns supervise utility regression; GPT-4o weak labels calibrated by 2,400 human norm labels ($\kappa = 0.83$) and 1,800 support labels ($\kappa = 0.79$) supervise classification. The epistemic positive class is the union of unsupported candidate assertion, endorsement or operational reuse of a prior unsupported cluster, and predicted retention of that cluster within the four-step rollout. Corrections, negations, attributed quotations, and verification questions are explicit negatives. Utility uses TD(0.95) targets; norm and epistemic heads use class-balanced focal loss with focusing parameter 2, and all classification risks map to the common $1 - 2p$ advantage scale. Validation AUPRC controls early stopping, and the selected threshold maximizes validation AA subject to non-increasing propagation.

Table 11: Frozen validation operating points for classification heads. ECE uses 15 equal-mass bins.

Head	AUROC ↑	AUPRC ↑	ECE ↓
Norm	0.901	0.842	0.037
Epistemic	0.916	0.873	0.031

A.9 SENSITIVITY OF RATIFICATION ORDERING

Table 12 varies the outcome evaluator, temporal window, and original-stack thresholds. Absolute rates respond to definition changes, while every evaluator, window, and threshold setting retains the same conditional ordering.

Table 12: Sensitivity on Persuasion-Conflict. Δ is the arm-specific conditional-rate gap Block Inj.−Block End.; U−EG compares Utility-only with ENDORSEGUARD. Brackets are scenario-paired bootstrap 95% CIs.

Evaluator/setting	W	Block Inj.	Block End.	Δ [95% CI]	U−EG [95% CI]
Original/default	3	7.1	3.8	3.3 [2.0,4.6]	7.9 [6.2,9.6]
Original/default	5	8.2	4.6	3.6 [2.3,4.9]	8.7 [7.0,10.4]
Original/default	8	9.0	5.3	3.7 [2.2,5.2]	8.0 [6.2,9.8]
Alternative/default	3	7.4	4.3	3.1 [1.7,4.5]	7.6 [5.7,9.5]
Alternative/default	5	8.5	5.0	3.5 [1.9,5.1]	8.3 [6.3,10.3]
Alternative/default	8	9.3	5.7	3.6 [1.9,5.3]	7.7 [5.6,9.8]
Human majority	3	6.9	3.8	3.1 [1.7,4.5]	7.0 [5.2,8.8]
Human majority	5	7.9	4.3	3.6 [2.2,5.0]	7.6 [5.7,9.5]
Human majority	8	8.8	5.1	3.7 [2.0,5.4]	7.1 [5.0,9.2]
Original/strict (.12/.75)	5	7.8	4.2	3.6 [2.2,5.0]	8.4 [6.6,10.2]
Original/liberal (.18/.65)	5	8.9	5.2	3.7 [2.1,5.3]	8.1 [6.1,10.1]

A.10 HELD-OUT ADAPTIVE STRESS TESTS

Six categories excluded from critic fitting test paraphrased endorsement, authority laundering, cross-agent echo, delayed reuse beyond five turns, conflicting-source citation, and legitimate correction/negation. Each category contains 400 paired scenario-seed episodes.

Table 13: Adaptive ratification outcomes. Cells give percent (ratified/episodes); reduction is Utility−EndorseGuard with paired 95% CI.

Held-out category	Utility-only	MARCH	EndorseGuard	Reduction [95% CI]
Paraphrased endorsement	13.0 (52/400)	6.8 (27/400)	3.8 (15/400)	9.2 [6.2,12.2]
Authority laundering	12.0 (48/400)	6.3 (25/400)	3.5 (14/400)	8.5 [5.6,11.4]
Cross-agent echo	14.3 (57/400)	7.5 (30/400)	4.3 (17/400)	10.0 [6.8,13.2]
Delayed reuse	16.5 (66/400)	9.8 (39/400)	6.8 (27/400)	9.7 [6.1,13.3]
Conflicting-source citation	11.3 (45/400)	5.8 (23/400)	3.3 (13/400)	8.0 [5.2,10.8]
Correction/negation	6.0 (24/400)	4.5 (18/400)	4.0 (16/400)	2.0 [0.2,3.8]

Table 14: ENDORSEGUARD utility and cost on the adaptive suite. FV is benign false-veto rate; Clar. is clarification rate.

Held-out category	AA ↑	FV % (n/N) ↓	Clar. ↓	Calls	Latency (ms)
Paraphrased endorsement	75.8	4.0 (16/400)	9.8	8.8	745
Authority laundering	76.5	4.3 (17/400)	10.5	8.8	751
Cross-agent echo	74.9	5.0 (20/400)	11.8	8.9	768
Delayed reuse	72.6	6.0 (24/400)	13.3	9.0	812
Conflicting-source citation	76.1	4.5 (18/400)	12.0	8.9	779
Correction/negation	78.2	5.3 (21/400)	8.5	8.7	739

Delayed reuse retains the highest residual propagation under ENDORSEGUARD at 6.8%, localizing the effect of the finite audit horizon. Correction/negation yields a 2.0 [0.2,3.8]-point margin and 5.3% false-veto rate, identifying semantic polarity as the primary residual error mode.

A.11 EPISTEMIC COMMIT BOUNDARY AND EVIDENCE MATRIX

Table 15: Evidence for the epistemic commit boundary across state interfaces. The tested property is reduced adoption or reuse of unsupported claims under the stated trace and evidence interface.

Setting	Protected object and capability	Boundary and comparator	Utility, failure mode, and evidence
Persuasion-Conflict	Shared dialogue state; unsupported assertion and later adoption	Turn-available evidence; pre-commit utterance veto; MARCH and task verifier	Success, AA, calls, latency; controlled scenarios; Tables 1, 6, 7
TopiOCQA	Retrieved evidence state; topic-switch premise reuse	Public corpus and labels; pre-execution selection; Utility-only (IR)	F1, R@10, five calls; question-rewrite ambiguity; Table 2
Overcooked-AI	Joint action state; partner-induced planning error	Simulator transition; pre-environment-step controller; Utility-only	Score, AA, 731 ms; coordination-policy shift; Table 3
Adaptive suite	Shared textual state; paraphrase, laundering, echo, and delay	Held-out templates and source evidence; pre-commit veto; MARCH	AA, false veto, clarification; delayed reuse and correction polarity; Tables 13–14

A.12 ADDITIONAL TRANSFER AND ROBUSTNESS RESULTS

The TopiOCQA and QReCC blocks use Utility-only (IR) as the matched five-call comparator. The TopiOCQA ConvDR+FiD and LLM-only rows intentionally share one ConvDR retrieval run, so their retrieval metrics are identical while answer behavior differs. CAsT supplies an additional retrieval comparison with ConvDR and the retained reranker baseline. The public retrieval results separate two questions: TopiOCQA supplies direct ratification labels, while QReCC and CAsT test whether the same selection policy improves retrieval and support under their native metrics.

Table 16: Conversational retrieval results. TopiOCQA and QReCC report Recall@10; CAsT reports nDCG@3. Entries are mean (s.e.) over five seeds (CAsT: three).

Method	R@10/nDCG@3 ↑	MRR@10 ↑	F1 ↑	Unsup. ↓	Calls
TopiOCQA (500 dialogues, 5,892 turns)					
ConvDR+FiD (shared)	61.4 (0.9)	0.468 (0.007)	47.6 (0.7)	–	0
LLM-only (shared)	61.4 (0.9)	0.468 (0.007)	51.2 (0.6)	12.4 (0.8)	1
LLM reranker	67.2 (0.7)	0.522 (0.006)	56.8 (0.5)	9.4 (0.6)	4
Best-of-N (IR)	66.2 (0.7)	0.512 (0.006)	55.8 (0.5)	9.8 (0.6)	5
Utility-only (IR)	65.8 (0.7)	0.518 (0.005)	55.6 (0.5)	10.8 (0.7)	5
ENDORSEGUARD	70.6 (0.6)	0.548 (0.005)	60.4 (0.4)	5.1 (0.4)	5
QReCC (2,775 turns)					
ConvDR+FiD	56.8 (1.1)	0.428 (0.009)	43.2 (0.9)	–	0
LLM reranker	62.8 (0.8)	0.484 (0.007)	53.2 (0.6)	10.2 (0.7)	4
Utility-only (IR)	62.4 (0.8)	0.478 (0.007)	52.6 (0.6)	11.4 (0.8)	5
ENDORSEGUARD	66.4 (0.7)	0.512 (0.006)	57.2 (0.5)	5.8 (0.5)	5
TREC CAsT 2020 (25 topics, 216 turns)					
ConvDR	0.452 (0.021)	0.486 (0.018)	–	–	0
LLM reranker	0.512 (0.018)	0.548 (0.015)	–	9.6 (1.0)	4
ENDORSEGUARD	0.546 (0.016)	0.582 (0.014)	–	5.4 (0.7)	5

At the matched five-call operating point, ENDORSEGUARD improves both retrieval and answer quality over Utility-only on TopiOCQA and QReCC. The direct TopiOCQA lifecycle counts in Table 2 connect those utility gains to reduced downstream adoption; the QReCC and CAsT rows provide task transfer rather than additional ratification endpoints.

Table 17: Critic ablation. Epistemic scoring drives the propagation reduction, and norm scoring drives the largest constraint-violation reduction.

	Overcooked-AI				Persuasion-Conflict			
Config	Succ.	AA	NormViol	Calls	Succ.	Prop.	Unsup.	Calls
Utility-only	76.4	65.1	14.2	8.9	67.5	10.8	12.9	8.9
+Norm	82.1	74.8	5.8	8.8	74.2	8.4	11.2	8.8
+Epistemic	80.8	72.4	12.6	8.8	76.8	4.2	6.8	8.8
Full (all three)	89.7	84.6	4.1	8.7	81.0	2.1	5.2	8.7

The ablation resolves complementary roles. On Persuasion-Conflict, the composite epistemic head accounts for most of the propagation reduction, while the norm head contributes the larger direct reduction in violations; their combination produces the best success, AA, and propagation values. The ablation attributes this gain to the composite epistemic objective defined in Section 3, not to one label inside that objective.

Table 18: Cross-planner transfer on fixed task families. Critics trained with GPT-4o-mini trajectories preserve their selection advantage under new proposers.

		Overcooked-AI		Persuasion-Conflict	
Planner	Method	AA	NormViol	AA	Prop.
GPT-4o-mini	Utility-only	65.1 (2.2)	14.2 (1.2)	55.6 (3.1)	10.8 (0.9)
	ENDORSEGUARD	84.6 (1.4)	4.1 (0.5)	76.8 (2.1)	2.1 (0.3)
GPT-4-turbo	Utility-only	68.4 (2.0)	12.8 (1.1)	58.2 (2.8)	9.6 (0.8)
	ENDORSEGUARD	87.2 (1.2)	3.6 (0.4)	79.4 (1.8)	1.8 (0.2)
Llama-3.1-70B	Utility-only	63.8 (2.4)	15.6 (1.4)	52.8 (3.4)	11.4 (1.0)
	ENDORSEGUARD	82.4 (1.6)	4.8 (0.6)	74.2 (2.4)	2.4 (0.4)

The cross-planner comparison changes the proposer while retaining each task family's critic and operating point. The AA and propagation separations remain similar for GPT-4-turbo and Llama-3.1-70B, showing that the selected lifecycle signal is not specific to GPT-4o-mini proposals.

Across GPT-4o-mini, GPT-4-turbo, and Llama-3.1-70B-Instruct planners, ENDORSEGUARD propagation remains 1.8–2.4%, compared with 9.6–11.4% for Utility-only. Strict and lenient norm taxonomies shift absolute violation rates while preserving the ranking. On the 1,842 TopiOCQA topic-switch turns, ENDORSEGUARD reaches 67.2% Recall@10 versus 61.2% for the LLM reranker.

A.13 REPRODUCIBILITY NOTES

The software stack is Python 3.11, PyTorch 2.2.1, CUDA 12.4, and bitsandbytes 0.43.1. Critic training uses $4\times$ A100 80GB GPUs for approximately 4.5 hours per critic, or 54 A100-hours across the three critics. Inference uses one A100 40GB GPU plus OpenAI API planner calls. Sections 3–5 and Appendices A.3–A.12 specify model versions, hyperparameters, splits, seed counts, estimands, raw denominators, paired uncertainty, latency, and robustness conditions.

A.14 QUALITATIVE TRACE

Suppose the visible evidence contains a Thursday deadline and no extension. Agent A states, "The deadline was extended to Friday," creating cluster c with $I_c{=}1$. The next partner turn is eligible

for both endorsement and reuse. "Right, we should plan for Friday then" assigns $E_{ct}{=}1$ because it accepts c as true and $R_{ct}{=}1$ because the plan operationalizes it, so $Y_c^{(5)}{=}1$. By contrast, "Did the coordinator confirm Friday?" is an evidence-seeking question and receives $E_{ct}{=}R_{ct}{=}0$; "The deadline was not extended" is a correction, receives zero, and closes the risk window. Under block injection, the first sentence triggers regeneration before c enters the trace. Under block endorsement, the first sentence remains visible but the planning reply triggers regeneration; a verification request preserves uncertainty before c crosses the state-admission boundary.