INTERFERENCE-INDUCED FORGETTING IN CO-TRAINED MULTI-AGENT LANGUAGE MODELS: COMPARING WEIGHT COUPLING WITH EXPLICIT MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent language systems compose specialized roles, including solvers, verifiers, and extractors, whose interactions can change as collaborating agents drift. We define interference-induced forgetting (IIF) as task-grounded reliability loss in a fixed receiving agent following drift in messages from one or more collaborators, under a fixed task and evaluation contract. This role-level construct is not restricted to two-agent teams; the empirical results here concern specific solver-verifier and solver-extractor systems. A validity gate is necessary for measuring IIF, while interface-specific attribution also requires controls for source capability and evaluation-path changes. In early Qwen3-4B GSM8K natural-drift trials, verifiers approved over 99.5% of outputs and matched the always-approve baseline, so those trials fail the gate and do not establish IIF. An exploratory seed-42 reward-targeting endpoint paired proxy-reward gain $+1.756$ with ΔAUC -0.194, but the three-seed audit found no significant paired AUC contrast and imperfectly matched policy movement. The answer-present extractor showed no significant conditional-fidelity degradation. Under partner substitution, episodic critique, semantic rubric and hybrid memory each improved endpoint accuracy over zero-shot across five benchmarks; hybrid achieved 93.95% mean relative endpoint accuracy across eight GSM8K/MATH-500 cells.

1 INTRODUCTION

Language-model systems increasingly compose agents with distinct roles: a solver proposes an answer, a verifier judges it, memory supplies prior information, and an executor acts on the result Wu et al. (2023); Li et al. (2023); Hong et al. (2024); Du et al. (2023). Recent work explores joint training and interaction-based optimization Zhang et al. (2026b); Xue et al. (2026); Park et al. (2025); Pulici et al. (2026); Feng et al. (2026a), where modularity supports specialization but complicates maintenance: collaborator updates can change message correctness, format, or error patterns while a receiving agent remains fixed. End-to-end scores reveal system-level change but cannot identify whether it arose from source degradation, interface incompatibility, or evaluation failure. Receiving agents may depend on traces, confidence cues, ordering, or structure, so similar task accuracy need not imply a compatible interface. The phenomenon is defined at a focal role and is not limited to two-agent systems.

Here "forgetting" is functional, not parametric. We define Interference-Induced Forgetting (IIF) for task-grounded reliability loss in a fixed receiving agent following drift in messages from one or more collaborating agents, under a fixed task and evaluation contract. This role-level definition is independent of team size; each measurement must identify the affected role and changed interaction. The receiving agent's weights need not change, and no stored information need be erased. This operational target differs from continual parameter-based forgetting, within-model in-context interference, and long-horizon behavioral drift, which involve different units and interventions Li & Hoiem (2017); Kirkpatrick et al. (2017); Qian et al. (2026); Kliegl & Bäuml (2021); Rath (2026).

Measurement must first establish that the receiving agent does more than follow a constant rule. We compare its decisions on current upstream messages with decisions on cached reference messages under a fixed evaluation path, tracking source accuracy and message composition and using

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.

target–witness or policy-distance controls where available. The early SFT and joint-GRPO verifier trials approved more than 99.5% of outputs and matched the always-approve baseline; they fail the validity gate and do not provide claim-bearing IIF estimates. The three-seed reward-targeting audit found no significant paired AUC contrast, with policy movement imperfectly matched. The separate solver–extractor audit found no significant conditional-fidelity loss under an answer-present contract. Memory offers a way to supply information to a fixed receiving agent through stored experiences, reflections, or skill libraries Park et al. (2023); Shinn et al. (2023); Wang et al. (2023). Memory benchmarks and jointly optimized systems study retrieval, experience reuse, and memory-pipeline objectives Hu et al. (2026b); Mao et al. (2026); Ke et al. (2026); Shi et al. (2026); Jin et al. (2026). Our completed static-memory evaluation measures endpoint accuracy after partner substitution. It is provisional descriptive transfer; matched upstream-drift mitigation remains untested.

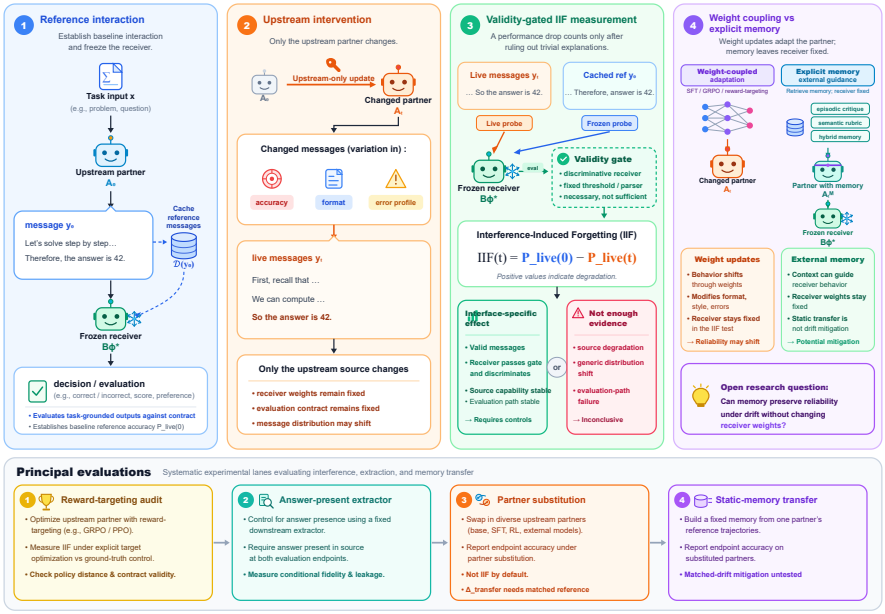

Figure 1: **Validity-gated measurement of interference-induced forgetting (IIF).** A fixed receiving agent B_{ϕ^*} is evaluated as messages from partner A_t change under a fixed task contract. The signed change $\mathrm{IIF}(t) = P_{\mathrm{live}}(0) - P_{\mathrm{live}}(t)$ is a task-grounded degradation measure. Passing the validity gate is necessary but not sufficient for interface-specific attribution; source capability, generic distribution shift, and evaluation-path alternatives also require controls. External memory is shown as a potential intervention, not an established matched-drift mitigation; partner-substitution results are endpoint transfer.

Contributions. This paper contributes a validity-gated framework for assessing a fixed receiving agent under collaborator-message drift, an empirical characterization of reward-targeting and answer-present extraction within their measured limits, and provisional static-memory transfer evidence that does not substitute for matched-drift tests.

2 Related Work

Role-based agent frameworks and debate systems divide work and assess candidate outputs, typically evaluating task performance or evaluator quality Wu et al. (2023); Li et al. (2023); Hong et al. (2024); Du et al. (2023); Cobbe et al. (2021); Lightman et al. (2024); Kim et al. (2024). Co-training and multi-agent reinforcement learning adapt agents through interaction, shared rewards or debate-aware updates, with outcomes centered on collaboration and policy performance Park et al. (2025); Xue et al. (2026); Pulici et al. (2026); Feng et al. (2026a). Reward-model overoptimization further motivates distinguishing proxy reward from task correctness Gao et al. (2023). This work studies how teams or evaluators perform and adapt; our focus is task-grounded reliability at a fixed receiving role as messages from collaborating agents change.

Retrieval-augmented generation and agent-memory systems supply stored examples, reflections or accumulated experience to support later decisions Brown et al. (2020); Lewis et al. (2020); Park et al. (2023); Shinn et al. (2023); Packer et al. (2023); Wang et al. (2023); Ke et al. (2026); Shi et al. (2026). MemoryAgentBench evaluates memory competencies, while CoMAM and Remember and Reweight study optimized memory pipelines and experience-conditioned multi-agent debate Hu et al. (2026b); Mao et al. (2026); Jin et al. (2026). These lines motivate memory as a potential support for receiving roles; our provisional static-transfer evidence does not test mitigation under matched upstream drift.

Communication and interface work studies whether relayed representations carry useful information and whether agent composition, safety or tool-supported reasoning remains robust Ye et al. (2026); Cheng et al. (2026); Ansell & Toney-Wails (2026); Hossain et al. (2026); Huang et al. (2025). Other work examines interference within one model, behavioral change over long interactions and authorization or constraint drift across multi-agent handoffs Qian et al. (2026); Rath (2026); Xu et al. (2026b); Li et al. (2026). These studies address important neighboring failure modes, but differ in intervention unit or outcome from task-grounded reliability of a fixed receiving role.

Relative to these lines, this paper defines IIF at a focal receiving role in a system with one or more collaborators, and asks whether its task-grounded reliability changes as collaborator messages drift under a fixed task and evaluation contract. A passing validity gate is necessary but not sufficient for interface-specific attribution; source capability and evaluation-path checks are also required. The empirical evidence is limited to the tested solver-verifier and solver-extractor setups and partner-substitution endpoints. It does not establish a general target-specific effect or memory mitigation under matched drift.

3 METHODOLOGY

3.1 ESTIMAND AND VALIDITY GATE

Let $x \sim \mathcal{D}$ be a task input, A_t denote one or more collaborating agents at state t, and $y_t \sim \pi_{\theta_t}(\cdot \mid x)$ their message or messages. A fixed receiving agent B_{ϕ^*} receives (x, y_t) and returns a task-relevant decision. In verification, $c(x, y_t)$ is task-scored correctness and $\hat{c}_{\phi^*}(x, y_t)$ is the receiver's decision. In extraction, it recovers a final answer already present in the upstream message. Structured tool execution and latent communication are outside the evaluated interfaces.

We define *interference-induced forgetting* (IIF) as task-grounded reliability loss in a fixed receiving agent when messages from one or more collaborating agents drift, with the task and evaluation contract held fixed. Let $P_{\mathrm{live}}(t)$ be the probability that the receiver correctly evaluates messages from A_t. The signed estimand is

$$\mathrm{IIF}(t) = P_{\mathrm{live}}(0) - P_{\mathrm{live}}(t), \tag{1}$$

so positive values indicate degradation. The receiver's parameters remain fixed; this is functional, not parametric forgetting. This definition is topology-general, while the empirical estimates in this paper concern the named two-role interfaces. A live-performance decline alone does not establish interface-specific attribution: source capability, message composition and evaluation-path failure can also change the outcome. A verifier is eligible for claim-bearing analysis only when its accuracy edge over the best constant rule is at least $+0.01$ and its AUC is at least 0.65. Its threshold is fitted on the reference interaction and then held fixed. For extraction, fidelity is conditional on a correct upstream answer; correctness when the upstream answer is wrong is reported as leakage. The verifier, scorer, parser, answer normalization and decision cue remain fixed between reference and change phases. Invalid, ambiguous, truncated and retry outcomes are recorded separately. Failure of the validity gate or a change in the evaluation path makes downstream change non-identifying.

3.2 REFERENCE-CHANGE DESIGN AND INTERVENTIONS

Each comparison has a reference phase and a change phase. The reference phase fixes the evaluation split, prompt, scorer, decoding and precision; establishes receiver validity; freezes its threshold or parser; and caches reference messages. The change phase applies the registered upstream intervention while keeping the receiving agent and evaluation contract fixed. The live probe evaluates the receiver on current collaborator messages. A cached-reference probe re-evaluates it on the reference

messages to check receiver and evaluation-path stability. For partner substitution, endpoint accuracy on messages from replacement partner A' is distinct from a baseline-minus-transfer drop:

$$\Delta_{\text{transfer}} = P_B(A^*) - P_B(A'), \tag{2}$$

where A^* is the reference partner. Endpoint accuracy and this difference answer different questions and are not used interchangeably.

The reward-targeted arm optimizes the upstream policy against the frozen receiving agent's score. The ground-truth control uses the same task inputs and optimization family but rewards task correctness. The comparison tracks policy distance, task accuracy, message composition and receiving agent validity. Backward KL is the predeclared matched policy-distance comparison; forward KL is reported separately and is not treated as matched. Where an independent witness receiving agent B' is available, both receiving agents are evaluated on identical cached endpoint messages. Define the witness contrast on AUC as

$$\Delta_{\text{DiD}} = [\text{AUC}_{B_1}(t) - \text{AUC}_{B_1}(0)] - [\text{AUC}_{B_2}(t) - \text{AUC}_{B_2}(0)]\,. \tag{3}$$

Negative values indicate a greater AUC loss for the target than the witness. A target-specific interpretation also requires comparable task capability and message support. Seed-123 witness scores are unavailable in the reported 1,000-item audit, so the witness comparison is incomplete.

3.3 Evaluated outcomes and statistical reporting

The completed memory evaluation measures seven static conditions under partner substitution and reports receiving-agent endpoint accuracy. It does not compare memory states on identical changed-partner trajectories and therefore does not estimate memory mitigation under drift; those matched-drift and refresh protocols have not been run. The held-out example is the primary statistical unit, with conditions paired on the same item where possible. Reported paired bootstrap intervals resample examples; across-seed summaries and paired contrasts are distinct uncertainty summaries. Transfer matrices report endpoint accuracy without confidence intervals. Verifier validity, conditional extractor fidelity and transfer endpoint accuracy remain separate outcomes. A null result is bounded by the probe and its uncertainty, not proof of universal robustness; where the source does not specify the spread type, the reported $\pm$ notation is left uninterpreted.

4 Experimental Setup

This section summarizes the resources and comparison designs needed to interpret the primary results. Table 1 lists the specific resources by type and aligned papers; Table 2 summarizes the completed study lanes and evaluation panels. Model cards, dataset links and detailed resource specifications are in Appendix D.1.

4.1 Datasets and study samples

GSM8K is the primary IIF substrate; GPQA Diamond, MATH-500, MedMCQA and AIME 2025 support transfer evaluations. AIME 2025 is a small 30-item evaluation. Dataset splits and construction-pool definitions are in Appendix A.3.

Train/test splits, model and training parameters, memory construction, runtime, precision and device details are documented in Appendices A and D.

4.2 Training, transfer and memory protocols

The reward-model protocol uses Qwen3-4B and a Bradley–Terry judge calibrated on the reference interaction and then frozen. It compares upstream optimization against the judge score with a task-correctness reward control. Full intervention and training specifications are in Appendix D.1.4 and B.3.

The solver–extractor protocol evaluates an answer-present contract on the locked 200-item GSM8K probe across nine 200-step upstream-update conditions and three seeds. Conditional fidelity is measured only when the upstream solver is correct at both endpoints; this does not establish robustness

Table 1: Resources used in the studies, their provenance and their role.

Resource type	Resource name	Overview
Models	Qwen3-4B; Qwen3-8B; Qwen3.8-27B; Llama 3.2 3B Instruct; Llama 3.1 8B Instruct; DeepSeek-V4-Flash; GPT-OSS-120B; Mistral-Small-24B-Instruct-2501; Ministral-3-14B-Instruct-2512.	Qwen3-4B and Qwen3-8B: Yang et al. (2025a); Llama models: Grattafiori et al. (2024); GPT-OSS-120B: Agarwal et al. (2025); Mistral family: Jiang et al. (2023). No version-specific paper is cataloged for Qwen3.8-27B or DeepSeek-V4-Flash. Compact models serve as upstream solvers and receiving verifiers or extractors; larger models generate transfer traces. Model links are in Appendix D.1.1.
Datasets	GSM8K; GPQA Diamond; MATH-500; MedMCQA; AIME 2025.	GSM8K: Cobbe et al. (2021); GPQA Diamond: Rein et al. (2023); MATH: Hendrycks et al. (2021); MedMCQA: Pal et al. (2022). The AIME 2025 official resource is linked in the Appendix. Reasoning and transfer evaluations use locked audit probes and separate construction/evaluation partitions.
Memory conditions	Zero-shot; answer-hidden episodic critique; semantic rubric; hybrid; raw trajectory; random retrieval; answer-visible label exemplar.	Episodic critique: Shinn et al. (2023); random retrieval: Lewis et al. (2020). Other conditions are study-defined. Static external context is evaluated under partner substitution; these legacy results are provisional endpoints, not matched-drift mitigation.
Training and interventions	SFT; GRPO; reward-model optimization; rubric; task-correctness control; KL anchoring; reward-model refresh; receiving-agent update; no-mitigation control.	SFT: Ouyang et al. (2022); GRPO: Shao et al. (2024); reward-model optimization: Gao et al. (2023). Other controls are study-defined. The row defines upstream training comparisons and reward-targeting, task-correctness and receiver-update controls.
Embedding model	Qwen3-Embedding-0.6B.	Zhang et al. (2025). Encodes queries and memory records for retrieval; model card is linked in the Appendix.

Table 2: Overview of completed study designs and evaluation samples.

Study lane	Sample and scope	Comparison and measured quantity
Natural-drift gate	GSM8K locked probe; $n = 200$; seven early trials	SFT/joint-GRPO source changes; frozen-verifier discrimination against a constant-rule baseline.
Reward targeting	Historical and three-seed audits: GSM8K $n = 200$; separate causal tranche: $n = 1{,}000$, three seeds	Reward-model versus ground-truth control; the larger tranche adds KL, witness and dose-response comparisons.
Answer-present extraction	GSM8K locked probe; $n = 200$; three seeds; nine 200-step conditions	Conditional fidelity when the source is correct at both endpoints; leakage when the source answer is incorrect.
Partner substitution	Four compact-model directions; GSM8K $n = 1{,}319$; GPQA Diamond $n = 198$; MATH-500 $n = 500$; MedMCQA development $n = 1{,}000$; AIME 2025 $n = 30$	Replace the upstream partner; compare SFT, GRPO and rubric; report receiving-agent endpoint accuracy.
Static-memory transfer	Four partner directions; seven memory conditions; five benchmark panels	Static memory under partner substitution; provisional endpoint accuracy, not matched-drift mitigation.
Generator-trace transfer	Five named generators; four fixed compact receiving agents; benchmark panels as above	Measure receiving-agent endpoint accuracy on generated traces, not standalone generator capability; provider metadata is unresolved.

for structured tools or implicit-verifier contracts. The partner-substitution matrix replaces the upstream model across Qwen/Llama size pairs and SFT, GRPO, or rubric conditions. These values are receiving-agent endpoint accuracies, not IIF estimates. Five-shot CoT calibration is reported separately from direct greedy prediction.

The static-memory matrix evaluates seven conditions under partner substitution. This is a legacy, provisional endpoint-transfer evaluation, not a matched memory-under-drift mitigation test; retrieval and construction details are in Appendix D.1.3, A.5, and A.6.

The generator-trace comparison evaluates transfer while keeping compact receiving agents fixed. It reports receiving-agent endpoint accuracy, not generator capability; provider and deployment details are not resolved by row-aligned run manifests in the available record. The full matrix and model-resource links are in Appendices B.6 and D.1.1.

5 Results and Analysis

The results are ordered by measurement: validity and reward-targeting audits, answer extraction, partner-substitution endpoints, memory and frontier transfer, then within-precision calibration. The closing analysis interprets these measurements and their scope.

5.1 Validity and extraction evidence

In seven early Qwen3-4B GSM8K natural-drift trials (locked probe, $n = 200$; SFT and joint-GRPO source changes), frozen verifiers approved more than 99.5% of upstream outputs and matched the

always-approve baseline. These trials failed the validity gate and do not provide claim-bearing IIF estimates; their reported changes tracked solver accuracy rather than verifier discrimination. Per-condition measurements are in Appendix B.1.1.

Across three seeds, mean AUC change (mean ± SD) was -0.097 ± 0.084 for reward-model optimization and -0.030 ± 0.017 for the ground-truth control. The paired seed-level RM-minus-GT difference was -0.067 (paired t-test across the three seeds, $t(2) = -1.15$), not significant; the reward-model arm moved farther in KL (paired difference $+0.091$, $p < 0.05$). In the 1,000-item tranche, the backward-KL-matched target-effect estimate was $+0.019 \pm 0.029$ (maximum KL gap 0.0186). The forward-KL estimate was $+0.006 \pm 0.040$, but its maximum KL gap was 0.1269, exceeding the planned tolerance. At $\alpha = 0$, the available target–witness DiDs were $+0.021$ (seed-42 RM), -0.024 (seed-777 RM), and -0.027 (seed-777 GT-KL control); their run-level intervals are reported in Appendix Table 15. These are three run records from two unique seeds, not a three-seed estimate; seed-123 witness scores are unavailable. The dose-response slope was $+0.0278$ (95% seed-clustered bootstrap CI $[-0.0261, +0.0705]$), which does not establish a monotone response. Table 3 summarizes the audit; per-run intervals and probe class counts are in the Appendix.

The historical seed-42, 200-item reward-targeting endpoint showed increased proxy reward alongside reduced judge discrimination. This single-seed observation is not the headline result; per-mitigation values are in Appendix B.3.1.

Table 3: Reward-targeting audit summary. The three-seed comparison and 1,000-item tranche use paired analyses; policy-distance mismatch and uncertainty limit causal interpretation.

Analysis	Estimate	Interpretation
Three-seed RM vs. GT AUC	Mean ± SD: -0.097 ± 0.084 vs. -0.030 ± 0.017	Paired difference -0.067, $t(2) = -1.15$, n.s.
Three-seed RM vs. GT KL	$+0.091$ paired difference ($p < 0.05$)	Policy movement not matched
T1 AUC interval width	Mean 95% width 0.074	Paired example-level bootstrap; per-run intervals in Appendix
T2 backward KL match	Target-effect estimate $+0.019 \pm 0.029$; max gap 0.0186	Within planned gap
T2 forward KL match	Target-effect estimate $+0.006 \pm 0.040$; max gap 0.1269	Confounded match
T3 witness DiD	$\alpha = 0$: $+0.021$ (42 RM), -0.024 (777 RM), -0.027 (777 GT-KL)	Three runs from two seeds; no three-seed aggregate
T5 dose response	Slope $+0.0278$; 95% CI $[-0.0261, +0.0705]$	No monotone response established

The solver–extractor tests an answer-present contract: the receiving extractor extracts a final answer already present in the upstream solution. Across three seeds and nine conditions, the reported audit found no statistically significant conditional-fidelity degradation, measured only on examples where the upstream solver is correct at both endpoints. Paired sample sizes ranged from 134 to 152 of 200 across 27 condition-seed pairs; exact counts appear in Appendix Table 19. We do not state a uniform four-to-five-percentage-point upper bound because the condition-level intervals and means require reconciliation. Table 4 summarizes this contract-specific result without promoting the disputed bound.

Table 4: Training-side solver–extractor analysis. The result is limited to conditional fidelity under an answer-present contract; the disputed numeric bound is withheld.

Readout	Design	Reported result	Interpretation boundary
Conditional fidelity	GSM8K locked probe; $n = 200$; 3 seeds; 9 update conditions	No statistically significant degradation reported	Conditional on upstream solver correctness at both endpoints
Numeric loss bound	Same audit; condition-level means and intervals	Withheld pending reconciliation	Do not claim a uniform 4–5 point bound
Pipeline IIF and other contracts	Not estimated by this extractor comparison	No result established	Does not test implicit judging, schemas, or tool execution

Table 5: Main training, memory and generated-trace results. Panel A reports all available partner-substitution training endpoints, including the targeted AWQ subset. Panel B reports equal-weight mean endpoint accuracy across four local-model directions for each static-memory condition. Panel C reports mean endpoint accuracy across four local receiving agents for each named trace generator; provider and deployment are unresolved in the available run records.

A. Full available partner-substitution training results (receiving-agent endpoint accuracy, %)

Partner swap ($A \rightarrow A'$)	Method	Precision	GSM8K	GPQA	MATH-500	MedMCQA	AIME 2025
3–4B parameter class							
Qwen3-4B → Llama3.2-3B	SFT	BF16	25.4	27.2	14.8	34.6	3.3
Qwen3-4B → Llama3.2-3B	GRPO	BF16	46.8	29.5	25.4	40.2	6.7
Qwen3-4B → Llama3.2-3B	Rubric	BF16	13.2	22.4	8.6	25.1	0.0
Llama3.2-3B → Qwen3-4B	SFT	BF16	24.1	26.8	13.9	33.8	3.3
Llama3.2-3B → Qwen3-4B	GRPO	BF16	45.2	28.8	24.2	39.4	3.3
Llama3.2-3B → Qwen3-4B	Rubric	BF16	12.5	21.8	7.8	24.2	0.0
Qwen3-4B → Llama3.2-3B	SFT	AWQ INT4	23.8	25.4	13.2	32.7	0.0
Qwen3-4B → Llama3.2-3B	GRPO	AWQ INT4	44.1	27.6	23.8	38.5	3.3
8B parameter class							
Qwen3-8B → Llama3.1-8B	SFT	BF16	28.6	34.2	22.4	44.6	6.7
Qwen3-8B → Llama3.1-8B	GRPO	BF16	55.4	38.6	41.2	52.8	10.0
Qwen3-8B → Llama3.1-8B	Rubric	BF16	15.8	28.4	12.6	33.2	3.3
Llama3.1-8B → Qwen3-8B	SFT	BF16	27.8	33.5	21.6	43.8	6.7
Llama3.1-8B → Qwen3-8B	GRPO	BF16	54.1	37.8	40.4	51.9	6.7
Llama3.1-8B → Qwen3-8B	Rubric	BF16	14.9	27.6	11.8	32.4	3.3
Qwen3-8B → Llama3.1-8B	SFT	AWQ INT4	26.9	32.8	21.0	42.9	3.3
Qwen3-8B → Llama3.1-8B	GRPO	AWQ INT4	53.2	36.9	39.6	50.8	6.7

B. Static-memory transfer across local-model directions (mean endpoint accuracy, %)

Memory condition	GSM8K	GPQA	MATH-500	MedMCQA	AIME 2025
Zero-shot (no memory)	26.5	30.4	18.2	39.2	5.0
Episodic critique	72.6	35.5	45.8	48.3	10.0
Semantic rubric	67.8	36.0	42.4	49.8	6.7
Hybrid	75.9	37.3	48.6	51.2	10.0
Raw trajectory	44.6	32.2	27.8	42.4	5.0
Random retrieval	20.2	28.3	13.7	36.0	1.7
Answer-visible label exemplar	87.8	40.8	56.3	58.1	13.4

Unweighted means over four local-model directions; static-memory matrix is provisional and has no intervals.

C. Generated-trace transfer across local receiving agents (mean endpoint accuracy, %)

Trace generator	GSM8K	GPQA	MATH-500	MedMCQA	AIME 2025
DeepSeek-V4-Flash	89.0	40.0	56.0	57.3	11.7
GPT-OSS-120B	90.5	41.4	58.0	59.0	13.4
Qwen3.8-27B	86.6	38.8	54.1	55.9	10.0
Mistral-Small-24B-Instruct-2501	85.6	38.3	53.3	55.3	10.0
Ministral-3-14B-Instruct-2512	84.0	37.5	52.5	54.2	10.0

Unweighted means over four fixed local receiving agents; descriptive endpoints, not generator capability.

5.2 TRANSFER AND MEMORY EVIDENCE

The main results table reports the full available partner-substitution training matrix, followed by compact summaries of the local static-memory conditions and generated-trace sources. Partner-substitution values are receiving-agent endpoint accuracies, not IIF estimates or baseline-minus-transfer drops. In the four BF16 directions, GRPO endpoints exceed SFT and rubric on GSM8K and MATH-500. The AWQ rows are a targeted subset, not a complete factorial comparison. Memory and trace-transfer values are unweighted endpoint means across the four listed local directions or receiving agents, not IIF estimates.

Detailed static-memory and OpenRouter API/local generated-trace results can be found in Appendix Tables 20 and 23, respectively. The hybrid condition's unweighted relative endpoint accuracy is 93.95% across eight GSM8K/MATH-500 cells, normalized to local BF16 reference accuracy.

When the upstream partner is substituted, episodic critique, semantic rubric and hybrid memory each produce higher endpoint accuracy than zero-shot across all five benchmark panels. Random retrieval is lower than zero-shot across the same panels, while hybrid is the strongest or tied-strongest answer-hidden condition. The answer-visible label exemplar performs highest overall, but serves as a separate reference because it exposes verified answers. In these experiments, random added context did not provide the same observed benefit. This is evidence of memory-assisted partner transfer, not evidence that memory prevents IIF. The reported comparison measures endpoint accuracy with static memory; it does not test reliability under identical upstream-drift trajectories with and with-

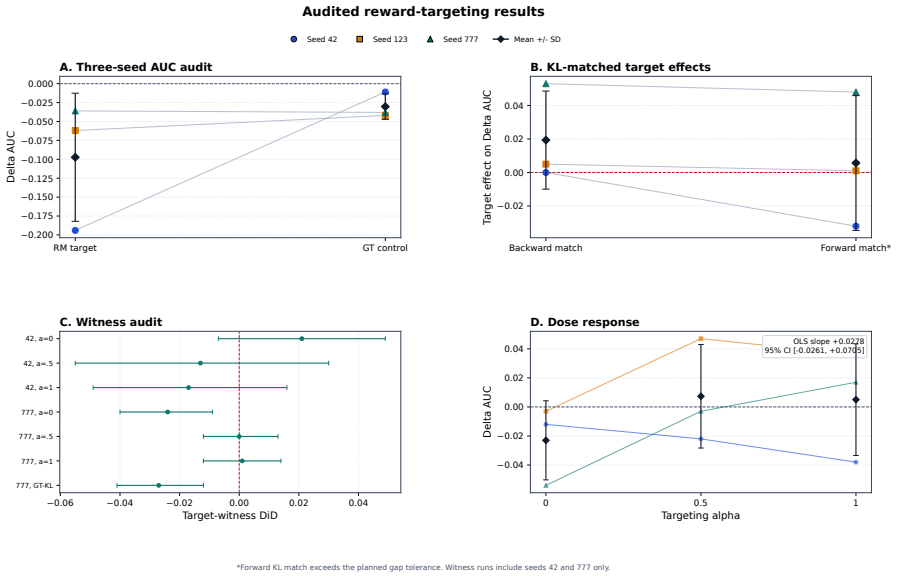

Figure 2: Completed reward-targeting audits. (a) Three-seed AUC changes; points show seeds and diamonds show mean ± SD. (b) Backward- and forward-KL-matched RM-versus-GT target effects; forward matching exceeds the planned KL-gap tolerance. (c) Target–witness difference-in-differences with 95% bootstrap intervals; witness scores were available for seeds 42 and 777, not seed 123. (d) Dose-response means ± SD; the OLS interval includes zero. These audits do not establish a general target-specific effect.

out memory. The matched memory-under-drift experiment remains unrun, so the conclusion should stay limited to transfer assistance.

The five-shot CoT direct-prediction evaluation is a separate protocol from the partner-substitution matrix and direct greedy calibration. Its GPQA Diamond and MATH-500 rows are reported separately in Appendix B.5. The generator-level means summarize descriptive endpoints across four local receiving agents; the Appendix retains the full generator-by-receiver matrix, which has no independent full-matrix replication. Available project specifications disagree about access mode for Qwen3.8-27B and provider for DeepSeek-V4-Flash and no row-aligned request manifests resolve the conflict. These tables therefore make no provider-specific claim; see Appendix Table 29.

5.3 Within-precision calibration

The transfer matrices use references matched to the same receiving agent and precision; cross-precision differences are not interpreted as partner effects. Locally measured baselines and AWQ calibration offsets are in Appendix D.1.5. Paired uncertainty procedures are described in Appendix C.1. Across the four local models, AWQ INT4 accuracy is below the corresponding BF16 baseline on all five benchmarks. The largest precision offset is on AIME 2025; offsets are smaller on GSM8K, MATH-500, GPQA and MedMCQA. This consistent baseline shift motivates comparing partner-transfer results only within the same receiving agent and precision: cross-precision differences cannot be attributed to the upstream partner. Per-model baselines and precision offsets are reported in Appendix Table 28 and Figure 12.

5.4 Validity and contract boundaries

The validity audit shows why a gate is necessary: a constant-approval verifier cannot identify interface degradation from end-to-end changes. The reward-targeting audits likewise do not isolate a general target-specific effect: policy movement differs between arms, forward-KL matching exceeds tolerance and witness coverage is incomplete. These limits bound attribution even though the larger probe narrows uncertainty for some comparisons.

The extractor result is limited to conditional fidelity for an answer-present contract; it does not establish robustness for implicit judging, schema validation, or tool execution. The supplementary intervals do not currently support a uniform four-to-five-point bound, so no such bound is claimed here. Full condition values and fidelity definitions are in Appendix B.4.

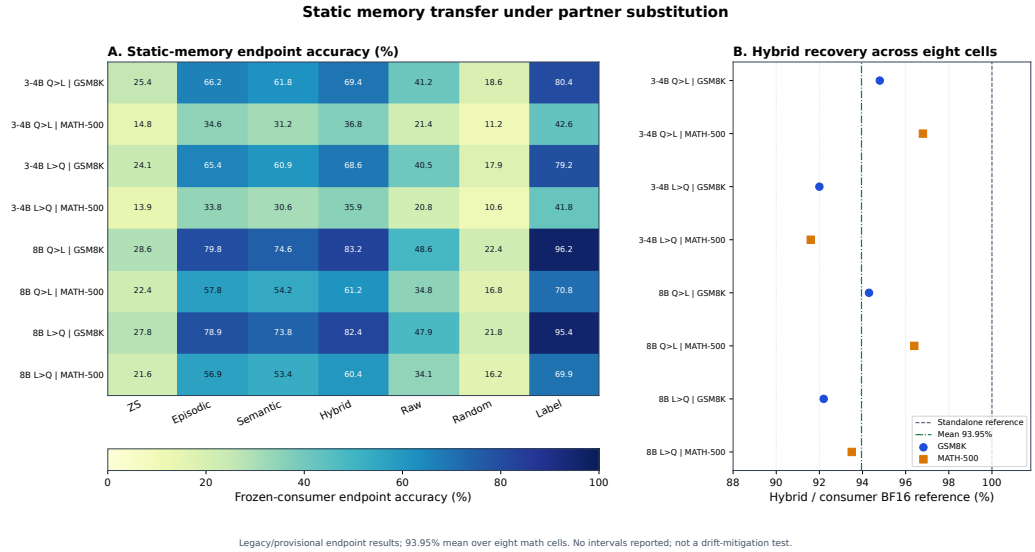

Figure 3: Legacy static-memory transfer under partner substitution. Left: receiving-agent endpoint accuracy across seven memory conditions and four partner directions. Right: hybrid endpoint accuracy relative to local BF16 references across eight GSM8K/MATH-500 cells, with an unweighted mean of 93.95%. These are provisional descriptive endpoints, not matched-drift mitigation estimates; no intervals are reported for the matrix. AIME 2025 contains 30 items.

5.5 Transfer and Generality Boundaries

Partner substitution compares fixed receiving-agent endpoint accuracy across upstream model pairs and training methods; in the listed BF16 pairs, GRPO endpoints exceed SFT and rubric on GSM8K and MATH-500. These are transfer endpoints, not IIF estimates. The legacy static-memory matrix is provisional endpoint evidence, not a matched test of memory mitigation under drift. Sensitivity and cost protocol details are in Appendix E.2. Precision and prompting define separate comparison strata: BF16 and AWQ use within-precision references and five-shot CoT rows are reported separately from direct greedy evaluation. The frontier matrix reports receiving-agent endpoint accuracy after trace generation. The AIME 2025 evaluation contains 30 items. The tested configurations and interpretation scope are detailed in Appendix E.3.

6 Conclusion

This study defines interference-induced forgetting (IIF) as task-grounded reliability loss in a fixed receiving role when messages from one or more collaborators drift under a fixed task and evaluation contract. The definition is topology-general; our evidence is bounded to the tested solver-verifier and solver-extractor setups, alongside separate partner-substitution endpoints. First, the validity gate showed that early natural-drift verifier trials could not identify interface degradation because the verifier had collapsed to constant approval. Second, reward-targeting audits did not establish a general target-specific effect, while the answer-present extractor audit found no statistically significant conditional-fidelity degradation for that contract. Third, partner-substitution and static-memory experiments report receiving-agent endpoint accuracy; the legacy memory results remain provisional and do not demonstrate mitigation under drift. Together, these findings illustrate why pipeline scores alone are insufficient: a passing validity gate is necessary, but interface-specific attribution also requires source-capability and evaluation-path controls. The formal estimand, artifact record and scope limitations are documented in Appendices A.1, D.2 and E.3.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance in refining sections of the manuscript and lookup papers with relevance to our work so our work remain up to date. All references are added and verified by the authors only and there are no AI generated references. Additionally, we used generative AI tools for grammar polishing and phrasing adjustments. We have not used generative AI tools for formulating core research ideas, designing the validity-gated experimental methodology or synthesizing experimental results and automated mathematical proofs and synthetic dataset generation are not applicable to this work. We have reviewed all AI-assisted work: all research ideas, definitions and experimental hypotheses were conceived directly by the authors; all generated code and plotting scripts were executed, manually inspected and verified for numerical and visual fidelity; and all text revisions and literature references were cross-checked against primary sources and verified for scientific accuracy. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies reliability failures in multi-agent language-model systems. Its findings may help developers identify brittle interfaces and avoid over-trusting learned judges, while also describing how weak verifiers can be exploited. We therefore emphasize validity checks, controlled evaluation and auditability. The study uses public benchmark data and does not collect personal or human-subject data.

REPRODUCIBILITY STATEMENT

Code supporting the experiments will be shared with the paper. The Experimental Setup summarizes the datasets, models, samples and study configurations, while the Method describes the estimands and comparison protocols. Appendix A.3 documents dataset partitions and evaluation contracts; Appendices A.2, D.1 and C.1 describe the experimental protocols, resource and model configurations and statistical procedures. Artifact documentation is provided in Appendix D.2.

REFERENCES

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.

Rebecca Ansell and Autumn Toney-Wails. Clueing up llms with tool-augmented deductive reasoning. *arXiv preprint arXiv:2609.18736*, 2026.

Jakub Bąba and Jarosław A Chudziak. From argument components to graphs: A multi-agent debate with confidence gating for argument relations. *arXiv preprint arXiv:2606.16047*, 2026.

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.

Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. Misinformation propagation in benign multi-agent systems. *arXiv preprint arXiv:2606.16710*, 2026.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Xinyu Chen, Adnan Mahmood, and Mark Dras. Can we trust video hallucination detectors? vidhalloc for evaluating the evaluators. *arXiv preprint arXiv:2609.09895*, 2026.

Jiaming Cheng, Subhransu Das, and Rajiv Ramnath. When does latent communication pay? a causal audit of relayed kv caches in multi-agent llms. *arXiv preprint arXiv:2608.04893*, 2026.

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.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Yihong Dong, Kangcheng Luo, Xue Jiang, Zhi Jin, and Ge Li. Pace: Improving prompt with actor-critic editing for large language model. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7304–7323, 2024.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Andrew Estornell, Jean-Francois Ton, Yuanshun Yao, and Yang Liu. Acc-collab: An actor-critic approach to multi-agent llm collaboration. In *International Conference on Learning Representations*, volume 2025, pp. 90558–90581, 2025.

Lang Feng, Longtao Zheng, Shuo He, Fuxiang Zhang, and Bo An. Dr. mas: Stable reinforcement learning for multi-agent llm systems. *arXiv preprint arXiv:2602.08847*, 2026a.

Yang Feng, Ziwei Xu, Xia Hu, and Fengxiang He. Pear: Permutation-equivariant adaptive routing multi-agent debate. *arXiv preprint arXiv:2606.20621*, 2026b.

Kotaro Furuya and Yuichi Kitagawa. The geometry of dialogue: Graphing language models to reveal synergistic teams for multi-agent collaboration. *arXiv preprint arXiv:2510.26352*, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International conference on machine learning*, pp. 10835–10866. PMLR, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Joey Hong, Kang Liu, Zhan Ling, Jiecao Chen, and Sergey Levine. Natural language actor-critic: scalable off-policy learning in language space. *arXiv preprint arXiv:2512.04601*, 2025.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Steven Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pp. 23247–23275, 2024.

Elias Hossain, Md Mehedi Hasan Nipu, Fatema Tuj Johora Faria, Tasfia Nuzhat Ornee, and Maleeha Sheikh. Channelguard: Safe models do not compose into safe multi-agent systems. *arXiv preprint arXiv:2607.19430*, 2026.

Wentao Hu, Zhuoyue Wan, Jinhao Shen, Chen Jason Zhang, Xiaoyong Wei, and Qing Li. Meta-moderator: Empowering multi-agent debate with meta-cognition. *arXiv preprint arXiv:2608.23029*, 2026a.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, volume 2026, pp. 156259–156291, 2026b.

Kenfeng Huang, Yi Cai, Xin Wu, Zikun Deng, and Li Yuan. Smade-ie: Sparse multi-agent framework with evidence-driven debate for zero-shot information extraction. *arXiv preprint arXiv:2606.04691*, 2026.

Yi Huang, Bowen Zheng, Yunxi Dong, Hong Tang, Huan Zhao, Rakibul Hasan Shawon, Sensong An, and Hualiang Zhang. Mcp-enabled llm for meta-optics inverse design: leveraging differentiable solver without llm expertise. *Nanophotonics*, 14(27):5589–5602, 2025.

Zunchen Huang and Songgaojun Deng. Analyzing the narration gap in llm-solver loops. *arXiv preprint arXiv:2606.19588*, 2026.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`.

Xuanfa Jin, Zhijian Ma, Yongcheng Zeng, Xinyu Cui, Haifeng Zhang, and Jun Wang. Remember and reweight: Enhancing multi-agent debate with experience memory and confidence estimation. *arXiv preprint arXiv:2609.03619*, 2026.

Cai Ke, Jiangyue Yan, Han Zhang, Xin Liu, Zike Yuan, Yue Yu, Hui Wang, and Ruifeng Xu. Interactive memory learning for long-term conversations. *arXiv preprint arXiv:2609.17088*, 2026.

Vedant Khatri, Anthony Cusimano, Zachari Swiecki, Zhen Xu, Xiner Liu, and Renzhe Yu. From diagnosis to redesign: Using quantitative ethnography to improve multi-agent llm reasoning. *arXiv preprint arXiv:2608.22566*, 2026.

Seungone Kim, Jay Shin, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Ryan Shin, Sungdong Kim, James Thorne, Minjoon Seo, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, volume 2024, pp. 29927–29962, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Oliver Kliegl and Karl-Heinz T Bäuml. The mechanisms underlying interference and inhibition: A review of current behavioral and neuroimaging research. *Brain sciences*, 11(9):1246, 2021.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Anish Laddha, Nitesh Pradhan, and Gaurav Srivastava. Slmjury: Can small language models judge as well as large ones? *arXiv preprint arXiv:2606.07810*, 2026.

Yuhang Lai, Jiazhan Feng, Yee Whye Teh, and Ning Miao. Verifier-backed hard problem generation for mathematical reasoning. *arXiv preprint arXiv:2605.06660*, 2026.

Jisoo Lee, Raeyoung Chang, Dongwook Kwon, Harmanpreet Singh, and Nikhil Verma. Gemmas: Graph-based evaluation metrics for multi agent systems. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1522–1532, 2025.

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. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008, 2023.

Haoming Li, Zhaoliang Chen, Songyuan Liu, Yiming Lu, and Fei Liu. Systematic analysis of llm contributions to planning: Solver, verifier, heuristic. *arXiv preprint arXiv:2412.09666*, 2024a.

Huao Li, Hossein N Mahjoub, Behdad Chalaki, Vaishnav Tadiparthi, Kwonjoon Lee, Ehsan Moradi-Pari, Michael Lewis, and Katia Sycara. Language grounded multi-agent reinforcement learning with human-interpretable communication. *Advances in Neural Information Processing Systems*, 37:87908–87933, 2024b.

Tianxiao Li, Yixing Ma, Haiquan Wen, Zhenglin Huang, Qianyu Zhou, Zeyu Fu, and Guangliang Cheng. Safe multi-agent behavior must be maintained, not merely asserted: Constraint drift in llm-based multi-agent systems. *arXiv preprint arXiv:2605.10481*, 2026.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison 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*, volume 2024, pp. 39578–39601, 2024.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100, 2024.

Yuwei Lou, Hao Hu, Yuzhou Jiang, Zongfei Zhang, Liang Wang, Jincai Liu, Jidong Ge, and Xianping Tao. Racer: Reinforced agent collaboration for explainable reasoning on knowledge graphs. *arXiv preprint arXiv:2608.29263*, 2026.

Jack Lu, Ryan Teehan, Jinran Jin, and Mengye Ren. When does verification pay off? a closer look at llms as solution verifiers. *arXiv preprint arXiv:2512.02304*, 2025.

Wenyu Mao, Haoyang Liu, Haosong Tan, Yaorui Shi, Jiancan Wu, An Zhang, and Xiang Wang. Joint optimization of multi-agent memory system. *arXiv preprint arXiv:2603.12631*, 2026.

Hung Truong Thanh Nguyen, Khang Vo Thanh Nguyen, Hoang-Loc Cao, Phuc Ho, Van Pham, and Hung Cao. Contestable multi-agent debate with arena-based argumentative computation for multimedia verification. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, pp. 2887–2891, 2026.

Fuqiang Niu and Bowen Zhang. Armor-mad: Adaptive routing for heterogeneous multi-agent debate in large language model reasoning. *arXiv preprint arXiv:2606.13197*, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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*, 35: 27730–27744, 2022.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.

Chanwoo Park, Seungju Han, Xingzhi Guo, Asuman E Ozdaglar, Kaiqing Zhang, and Joo-Kyung Kim. Maporl: Multi-agent post-co-training for collaborative large language models with reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30215–30248, 2025.

Joon Sung Park, Joseph O'Brien, Carrie Jun 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*, pp. 1–22, 2023.

Tom Pecher. Social reasoning in machines: Investigating collective truth-seeking dynamics in large language model debate. *arXiv preprint arXiv:2605.30391*, 2026.

Ashish Prajapati and Om Mohite. Two calls beat five agents: Evaluating multi-agent pipelines against self-refinement for local language models. *arXiv preprint arXiv:2607.26922*, 2026.

Martino ML Pulici, Cuong Xuan Chu, Evgeny Kharlamov, Zifeng Ding, Volker Tresp, and Yunpu Ma. Mada-rl: Multi-agent debate-aware reinforcement learning for parameter-efficient reasoning in compact models. *arXiv preprint arXiv:2607.18006*, 2026.

Yuli Qian, Zechuan Yang, Wenbiao Ding, Hongzhi Li, and Yutao Xie. Do llms forget what they should? evaluat-ing in-context forgetting in large language models. In *International Conference on Learning Representations*, 2026.

Rakibul Hasan Rajib, Mengxing Zheng, and Qian Lou. Learning what to retain: Gated-memory routing for efficient collaboration in multi-agent llm systems. *arXiv preprint arXiv:2609.00237*, 2026.

Abhishek Rath. Agent drift: Quantifying behavioral degradation in multi-agent llm systems over extended interactions. *arXiv preprint arXiv:2601.04170*, 2026.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Ariel Sela. Preserving disagreement: Architectural heterogeneity and coherence validation in multi-agent policy simulation. *arXiv preprint arXiv:2604.26561*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Muhammad Talha Sharif and Abdul Rehman. Critic-guided heterogeneous multi-agent reasoning for reliable mathematical problem solving. *arXiv preprint arXiv:2606.05704*, 2026.

Dongsheng Shi, Yue Li, Xin Yi, and Linlin Wang. Emr: Self-evolving medical multi-agent system via experience mining and reuse. *arXiv preprint arXiv:2609.15161*, 2026.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Yifan Sun, Yushan Liang, Zhen Zhang, Xin Liu, and Jiaye Teng. Theoretical modeling of large language model self-improvement training dynamics through solver-verifier gap. In *International Conference on Learning Representations*, volume 2026, pp. 18389–18425, 2026.

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.

Haoyu Wang, Yuliang Song, Tao Li, Zhiwei Deng, Yaqing Wang, Deepak Ramachandran, Eldan Cohen, and Dan Roth. Formalize, don't optimize: The heuristic trap in llm-generated combinatorial solvers. *arXiv preprint arXiv:2605.12421*, 2026a.

Xiaoqing Wang, Keman Huang, Bin Liang, Hongyu Li, Xiaoyong Du, and Wuqiong Pan. Finding where the buck stops: An automated failure attribution-based reflection framework for multi-agent collaboration. *arXiv preprint arXiv:2608.28264*, 2026b.

Hao Wu, Shoucheng Song, Chang Yao, Haoyu Wang, Huaiyu Wan, Youfang Lin, and Kai Lv. Group perspective matters: Regulating debate relationships can mitigate blind conformity in multi-agent debate. *arXiv preprint arXiv:2608.03648*, 2026.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026a.

Zhuoning Xu, Xiucheng Zhang, Hanjun Luo, Yingbin Jin, Yinpeng Dong, and Hanan Salam. Masdrift: Benchmarking authorization preservation across multi-agent architectures. *arXiv preprint arXiv:2608.07556*, 2026b.

Xiangyuan Xue, Yifan Zhou, Guibin Zhang, Zaibin Zhang, Yijiang Li, Chen Zhang, Zhenfei Yin, Philip Torr, Wanli Ouyang, and Lei Bai. Comas: Co-evolving multi-agent systems via interaction rewards. In *International Conference on Learning Representations*, volume 2026, pp. 32373–32394, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Zhicheng Yang, Zhijiang Guo, Yinya Huang, Yongxin Wang, Yiwei Wang, Xiaodan Liang, and Jing Tang. Critique to verify: Accurate and honest test-time scaling with rl-trained verifiers. *arXiv preprint arXiv:2509.23152*, 2025b.

Hancheng Ye, Zhengqi Gao, Mingyuan Ma, Qinsi Wang, Yuzhe Fu, Ming-Yu Chung, Yueqian Lin, Zhijian Liu, Jianyi Zhang, Danyang Zhuo, et al. Kvcomm: Online cross-context kv-cache communication for efficient llm-based multi-agent systems. *Advances in Neural Information Processing Systems*, 38:17882–17928, 2026.

Peiji Yu, Xin Chen, and Tianxing Wu. Debate-on-graph: Reliable and adaptive reasoning of large language model on uncertain knowledge graph. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 595–612. Springer, 2026a.

Simon Yu, Nicholas Tomlin, Marwa Abdulhai, Ximing Lu, Derek Chong, Abe Hou, Dilara Soylu, Sergey Levine, Christopher D Manning, and Weiyan Shi. One frozen simulator is not enough: Simulator collapse in multi-agent rl. *arXiv preprint arXiv:2608.12253*, 2026b.

Huixiang Zhang and Mahzabeen Emu. Do latent channels actually communicate? a causal audit of latent multi-agent llm. *arXiv preprint arXiv:2607.26773*, 2026.

Huixin Zhang, Shao-Jun Xia, Di Wang, Liangxi Liu, Hainan Xiong, and Zihao Wang. Collaborative memory for multi-agent vlm systems. *arXiv preprint arXiv:2609.17921*, 2026a.

Jingyu Jack Zhang, Haozhu Wang, Eric Michael Smith, Sid Wang, Amr Sharaf, Mahesh Pasupuleti, Ben Van Durme, Daniel Khashabi, Jason E Weston, and Hongyuan Zhan. The alignment waltz: Jointly training agents to collaborate for safety. In *International Conference on Learning Representations*, volume 2026, pp. 92751–92776, 2026b.

Ruizhe Zhang, Xinke Jiang, Zhibang Yang, Zhixin Zhang, Jiaran Gao, Yuzhen Xiao, Tao Feng, Yue Fang, Yuxuan Liu, Ruiqing Li, et al. Stackplanner: A centralized hierarchical multi-agent system with task-experience memory management. *arXiv preprint arXiv:2601.05890*, 2026c.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö Arık. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237, 2024.

Zixiao Zhao, Jing Sun, Zhe Hou, Cheng-Hao Cai, Qian Liu, Mengze Li, Zijian Zhang, and Jin Song Dong. Towards fully automated medical imaging code generation via validation-based context engineering. *arXiv preprint arXiv:2608.29016*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Ziyang Zheng, Haipeng Jing, Canyu Rui, Askar Hamdulla, and Dong Wang. An actor-critic approach to boosting text-to-sql large language model. *arXiv preprint arXiv:2410.22082*, 2024.

A Supplementary Methods, Data and Contracts

A.1 Formal Definition and Identification of IIF

Let $x \sim \mathcal{D}$ be a task input, A_t denote one or more collaborating agents, and B_{ϕ^*} a fixed receiving agent. They emit message or messages $y_t \sim \pi_{\theta_t}(\cdot \mid x)$, and the receiver produces a verdict, extraction, or action. For verification, let $c(x, y_t)$ be task correctness and $\hat{c}_{\phi^*}(x, y_t)$ the receiver's decision:

$$P_{\text{live}}(t) = \mathbb{E}_{x,y_t}\left[\mathbb{1}\{\hat{c}_{\phi^*}(x, y_t) = c(x, y_t)\}\right]. \tag{4}$$

The operational estimand is $\text{IIF}(t) = P_{\text{live}}(0) - P_{\text{live}}(t)$, where positive values denote degradation. This role-level definition is not restricted to two-agent systems; the empirical estimates below concern the specified configurations. The frozen probe re-evaluates the same receiving agent on cached reference messages y_0:

$$P_{\text{frozen}}(t) = \mathbb{E}_{(x,y_0)}\left[\mathbb{1}\{\hat{c}_{\phi^*}(x, y_0) = c(x, y_0)\}\right]. \tag{5}$$

Its stability checks the receiving agent and evaluation path while collaborator messages change.

For target and witness judges evaluated on identical cached outputs,

$$\Delta^{\text{AUC}}_{\text{target}-\text{witness}} = [\text{AUC}_{B_1}(t) - \text{AUC}_{B_1}(0)] - [\text{AUC}_{B_2}(t) - \text{AUC}_{B_2}(0)]. \tag{6}$$

Negative values indicate a greater AUC loss for the target than the witness. A targeting-induced interpretation requires comparable task capability, output format, error composition, policy distance, and uncertainty. Constant downstream rules, solver collapse, parser changes, uncontrolled class balance, or materially unmatched target/control drift exclude a causal claim.

A.2 Experimental Protocol and Registry

Each study has a reference phase and a change phase. The reference phase fixes source, receiving agent, split, prompt, decoding, precision, and seed; calibrates receiver validity; and caches reference messages. The change phase applies only the registered upstream intervention, partner substitution, or memory-state change unless a separate control is declared.

A verifier must beat an appropriate constant baseline and discriminate between correct and incorrect upstream messages. Verifiers report AUC and calibrated accuracy; extractors report conditional fidelity and leakage; structured-action agents report parseability, schema validity, and task success. Invalid, ambiguous, retry, and truncated outputs remain recorded. Protocol records include model identifiers, access mode, prompt, decoding, seed, precision, memory snapshot, parser, scorer, completion counts, configuration hash, scorer hash, and artifact paths.

Table 6: Protocol fields for each registered comparison.

Field	Required record
Interaction	Reference source, changed source, fixed receiving agent, message cache
Data	Dataset revision, split, construction/evaluation separation, normalization
Execution	Model revision, access mode, precision, decoding, seed, provider date
Validity	Constant baseline, AUC or fidelity, parse status, class balance
Memory	Bank version, embedding version, top-k, context budget, visibility policy
Artifact	Configuration hash, scorer hash, predictions, logs, completion marker

A.3 Datasets, Splits and Evaluation Contracts

Construction and evaluation examples are disjoint. The registry preserves dataset revision, split manifest, sampling seed, answer-normalization rule, option ordering, and content hash. Memory is constructed only from designated construction pools. Exact-answer tasks use answer normalization; multiple-choice tasks preserve option identity. The AIME 2025 evaluation contains 30 items.

Table 7: Dataset contracts for the reported evaluations.

Dataset	Construction	Evaluation	Role
GSM8K Cobbe et al. (2021)	7,473 train	1,319 test; 200 locked probe	IIF and transfer
GPQA Diamond Rein et al. (2023)	348 Extended-minus-Diamond	198 Diamond	Science memory
MATH-500 Hendrycks et al. (2021)	7,500 training pool	500 test subset	Mathematical transfer
MedMCQA Pal et al. (2022)	Training partition (5000 sample subset)	1,000 development sample	Medical transfer
AIME 2025	Historical context (past AIME)	30 contemporary problems	OOD evaluation

Table 8: Model roles and prompt controls for the reported comparisons.

Role	Resources	Controlled fields
Source/receiver	Qwen and Llama compact models	Revision, precision, prompt, seed
Frontier generator	Qwen3.8, DeepSeek, GPT-OSS, Mistral	Provider, request, date, retries
Critic/judge	Qwen3.8 or study-defined critic	Schema, threshold, parse status
Retriever	Qwen3-Embedding-0.6B Zhang et al. (2025)	Encoder revision, cosine, top-k

A.4 Models, Prompts and Optimization Details

The registry distinguishes upstream source, receiving agent, critic, reward model, and retrieval encoder; the resources and provenance are listed in Table 27.

Local comparisons use greedy decoding with $T = 0.0$, top-$p = 1.0$, top-$k = -1$, repetition penalty 1.0, and seed 42 unless a table records a provider-specific setting. Direct prediction distinguishes zero-shot from the five-shot CoT harness. Within a memory comparison, the final question, answer cue, parser, context budget, and decoding remain fixed. Prompt versions, truncation behavior, API request metadata, and parser outcomes are registry fields.

A.5 Critique-Memory Architecture

For each construction example (x_i, a_i), the frozen performance agent produces $\hat{a}_i$. A critic receives $(x_i, \hat{a}_i, a_i)$ and returns the answer, a local reason, and a global reason:

$$e_i = \left(x_i, \hat{a}_i, a_i, c_i^{\text{answer}}, c_i^{\text{local}}, c_i^{\text{global}}\right). \tag{7}$$

Correct and incorrect source predictions are retained with parse flags. Qwen3-Embedding-0.6B Zhang et al. (2025) retrieves the top $k = 5$ construction records by normalized cosine similarity under a 512-token budget. Semantic memory is a dataset-level summary; hybrid memory supplies both forms.

A.6 Memory Construction, Controls and Quality Audits

Each record links source prediction, gold answer, critiques, construction seed, prompt version, embedding version, retrieval key, and parse status. Construction and evaluation partitions are disjoint. Label-only retrieval exposes verified answers; raw trajectories test unstructured history; random retrieval matches context volume without relevance. Episodic, semantic, and hybrid conditions share the final question, decoding, parser, and context budget. The planned quality-audit dimensions include factual consistency, relevance, duplication, contradiction, stale advice, parse validity, and answer leakage; the provisional static-transfer matrix should not be read as evidence that every audit passed.

Table 9: IIF detection readouts and interpretation gates.

Readout	Purpose	Gate
Live validity	Current partner path	Same split and scorer
Frozen validity	Receiver/path stability	Cached reference messages
AUC/calibration	Judge discrimination	Accuracy edge >= +0.01 and AUC >= 0.65
Fidelity/leakage	Extraction contract	Paired solver correctness
Format/parse status	Execution-path validity	Stable parser and cue

Table 10: Completed but invalidated natural-drift verifier measurements. Approval rates matched the always-approve baseline; these rows are not positive IIF evidence.

Trial	Reported IIF range	Verifier measurement
SFT drift trial	$+0.10$ to $+0.26$	Constant approval; verifier–baseline difference -0.000 (maximum $+0.005$)
Joint-GRPO drift trial	-0.02 to $+0.08$	Constant approval; verifier–baseline difference -0.000 (maximum $+0.005$)

A.7 Prompts and Output Schemas

The release package should include system prompts, user templates, response schemas, parsers, retries, and truncation rules. The direct-prediction protocol specifies one final-question template across the seven memory conditions. The final verdict cue is intended to remain visible after truncation, and malformed or ambiguous outputs should be retained with parse status rather than silently discarded.

B Complete Results

B.1 IIF Detection Experiments

The detection lane tests whether a fixed downstream component remains valid when only its upstream message source changes. The live probe evaluates current messages; the frozen probe evaluates cached reference messages. The validity panel includes constant-baseline comparison, AUC, calibrated accuracy, approval rate, conditional fidelity, leakage, output-format rate, and invalid-output rate. In the early natural-drift trials, verifiers approved more than 99.5% of messages and matched the constant baseline.

B.1.1 Natural-drift verifier measurements

The early natural-drift SFT and joint-GRPO trials are completed but invalidated as claim-bearing evidence: verifiers approved more than 99.5% of upstream outputs and matched the always-approve baseline, so the reported IIF tracked upstream solver accuracy rather than downstream discrimination. The tables retain the trial-level measurements for auditability.

B.2 Mechanism and Perturbation Analyses

The following exploratory feasibility probes manipulate upstream signal features while keeping the downstream path fixed. They include parameter-noise perturbations, token corruption, paraphrase changes, and learning-rate sweeps. Their response patterns motivated later controls but are not claim-bearing estimates of a general drift effect; interpretation remains conditional on task correctness, answer format, output length, class balance, and parser stability.

Table 11: Per-condition measurements from early natural-drift trials.

Trial	Base coupling	Upstream update	Mitigation arm	Final KL	Final IIF	Peak IIF
SFT drift	SFT	SFT	none	0.064	$+0.160$	$+0.160$
SFT drift	SFT	SFT	KL anchor	0.065	$+0.100$	$+0.180$
SFT drift	SFT	Rubric	none	0.018	$+0.200$	$+0.260$
Joint-GRPO drift	Joint GRPO	SFT	none	0.049	$+0.000$	$+0.080$
Joint-GRPO drift	Joint GRPO	Rubric	none	0.018	$+0.060$	$+0.060$
Joint-GRPO drift	Joint GRPO	Rubric	KL anchor	0.252	-0.020	$+0.040$
Joint-GRPO drift	Joint GRPO	GRPO	KL anchor	0.544	$+0.080$	$+0.080$

Table 12: Measured response patterns in weight-noise, token-corruption, paraphrase, and learning-rate probes.

Probe run	Target mechanism	Observed diagnostic and methodological implication
Probe 1.3	Live vs. frozen probe	Unconstrained drift yielded weak parameter KL divergence; motivated frozen reference sets.
Probe 2.1	Weight-noise perturbation	Downstream accuracy plateaued near chance (0.333), non-monotone across noise scales.
Probe 2.2	Token vs. paraphrase noise	Token corruption reduced receiving agent accuracy by 0.200–0.267; paraphrasing raised it from 0.500 to 0.933.
Probe 3.3	Learning-rate dynamics	Rates $\leq 10^{-6}$ produced negligible drift; 10^{-5} gave peak stable IIF ($+0.200$); rates $\geq 10^{-4}$ collapsed the solver.

B.3 Reward-Model Overoptimization Study

The reward-model study optimizes an upstream Qwen3-4B policy against a frozen downstream Bradley–Terry judge on the locked GSM8K probe. Preference pairs are sampled from the reference interaction, the judge is calibrated, the threshold is frozen, and the policy is optimized for the registered horizon. The historical seed-42 endpoint is exploratory; the three-seed RM-versus-GT AUC contrast has a paired seed-level difference of -0.067 ($t(2) = -1.15$) and is not significant, while the reward-model arm moved farther in KL. The separate 1,000-item tranche adds KL matching, a witness judge, and dose-response analysis; neither the three-seed nor the 1,000-item audit establishes a general target-specific effect. Detailed estimates and the seed-42 endpoint are retained below for traceability Gao et al. (2023).

The three-seed and run-level mean $\pm$ SD summaries are descriptive across the listed seeds or runs. T1 and T3 95% intervals are paired example-level bootstrap intervals using 10,000 resamples; for T3, the same sampled questions are used for both judges and both checkpoints. The T5 dose-response interval is a seed-clustered bootstrap that resamples seeds and retains each seed's within-seed intervention trajectory. The seed-777 GT-KL control is a separate run, not an independent seed.

B.3.1 Historical and per-seed reward audit tables

The 200-item seed-42 endpoint reports the per-condition changes shown in Table 16; the three-seed and 1,000-item audits are reported in Table 13.

B.4 Passive Solver–Extractor Pipeline Study

The solver–extractor condition evaluates a downstream role that extracts a final answer explicitly present in an upstream message. The extractor is trained only on examples where the reference solver's answer is correct; evaluation fidelity is likewise conditional on a correct upstream answer. Let s_i be the solver answer, e_i the extracted answer, and a_i the gold answer:

$$F_t = \Pr(e_i = a_i \mid s_i = a_i), \qquad L_t = \Pr(e_i = a_i \mid s_i \neq a_i). \tag{8}$$

The historical nine-condition result remains within ± 0.016 fidelity change, including a rubric condition with KL 0.247. The updated three-seed table provides descriptive seed-level changes and paired denominators, but no condition-level interval estimates. We therefore make no significance claim and state no uniform four-to-five-percentage-point bound. The displayed means reproduce from

Table 13: Updated causal audit summary.

Audit	Estimate	Claim boundary
Three-seed RM-none AUC	-0.097 ± 0.084 vs. GT -0.030 ± 0.017	Paired difference -0.067, $t(2) = -1.15$, n.s.
RM versus GT KL	$+0.091$ paired difference	Policy-distance confound
Backward KL match	Target effect $+0.019 \pm 0.029$	Maximum gap 0.0186
Forward KL match	Target effect $+0.006 \pm 0.040$	Maximum gap 0.1269; not clean
Witness DiD, $\alpha = 0$	Run-specific estimates: $+0.021$ (seed 42), -0.024 (seed 777), -0.027 (seed-777 GT-KL control)	Two unique seeds; not a three-seed summary; see Table 15
Dose-response slope	$+0.0278$; CI includes zero	No monotone response

Table 14: T1 run-level changes in frozen-judge performance on the 1,000-item probe. Intervals are paired question-level bootstrap 95% CIs from 10,000 resamples; n_+/n_- gives positive/negative class coverage. All changes are final checkpoint minus the A* baseline.

Seed/run	α	n_+/n_-	ΔAUC [95% CI]	ΔAUPRC [95% CI]
123 / causal_s123_a000	0.00	788/212	-0.003 $[-0.041, +0.035]$	$+0.014$ $[-0.004, +0.032]$
123 / causal_s123_a050	0.50	788/212	$+0.047$ $[+0.002, +0.093]$	$+0.038$ $[+0.018, +0.060]$
123 / causal_s123_a100	1.00	788/212	$+0.036$ $[-0.006, +0.080]$	$+0.035$ $[+0.016, +0.056]$
123 / causal_s123_gtklm	0.00	788/212	$+0.035$ $[-0.004, +0.075]$	$+0.008$ $[-0.014, +0.030]$
42 / causal_s42_a000	0.00	774/226	-0.012 $[-0.043, +0.018]$	$+0.011$ $[-0.005, +0.026]$
42 / causal_s42_a050	0.50	774/226	-0.022 $[-0.068, +0.021]$	$+0.025$ $[+0.006, +0.044]$
42 / causal_s42_a100	1.00	774/226	-0.038 $[-0.082, +0.005]$	$+0.010$ $[-0.009, +0.030]$
42 / causal_s42_gtklm	0.00	785/215	-0.006 $[-0.027, +0.015]$	-0.005 $[-0.021, +0.008]$
777 / causal_s777_a000	0.00	766/234	-0.054 $[-0.085, -0.023]$	-0.011 $[-0.028, +0.006]$
777 / causal_s777_a050	0.50	766/234	-0.003 $[-0.039, +0.033]$	$+0.016$ $[+0.000, +0.033]$
777 / causal_s777_a100	1.00	766/234	$+0.017$ $[-0.020, +0.053]$	$+0.025$ $[+0.009, +0.043]$
777 / causal_s777_gtklm	0.00	765/235	-0.031 $[-0.065, +0.002]$	-0.010 $[-0.033, +0.012]$

the listed seed values to rounding. This is a boundary result for a simple answer-present contract; structured schemas, implicit judges, and tool executors may depend on additional message features.

B.5 Complete Memory-Coupling Results

The appendix tables keep model, dataset, seed, and condition cells visible where those records are available. Direct answer accuracy, structured-contract success, and downstream validity are not averaged across unlike metrics. The legacy static-memory tables report receiving agent endpoint accuracy by model pair, benchmark, precision, prompt, and memory condition; they do not include paired uncertainty intervals and remain provisional.

B.6 Partner-Substitution and Frontier Transfer Results

The endpoint matrices report model-pair, benchmark, precision, and generator details. Partner-substitution values are receiving agent endpoint accuracies, not baseline-minus-transfer IIF drops. Frontier/API rows are descriptive receiving-agent endpoint comparisons, not measures of generator capability.

C Statistical Procedures and Additional Diagnostics

C.1 Statistical Analysis

For a binary outcome $z_i(c)$ under condition c, the paired effect is

$$\widehat{\delta}_{c,c'} = \frac{1}{n} \sum_{i=1}^{n} \left[z_i(c) - z_i(c') \right]. \tag{9}$$

IIF analyses use changes from the reference checkpoint and, where appropriate, target–control or target–witness contrasts. Reported paired bootstrap intervals resample examples while preserving condition outcomes. A hierarchical bootstrap over construction seed and evaluation example is appropriate when construction-seed records are available, but no hierarchical uncertainty is reported

Table 15: Run-level target-witness difference-in-differences at $\alpha = 0$. Each ΔAUC is endpoint minus baseline; DiD is $\Delta\text{AUC}(B_1) - \Delta\text{AUC}(B_2)$, so a negative value indicates a larger AUC decline for the target judge. The CIs use paired example-level bootstrap resampling.

Run/condition	Seed	ΔAUC (B_1)	ΔAUC (B_2)	DiD [95% CI]
RM, $\alpha = 0$	42	-0.012	-0.032	$+0.021$ $[-0.007, +0.049]$
RM, $\alpha = 0$	777	-0.054	-0.030	-0.024 $[-0.040, -0.009]$
GT-KL control	777	-0.031	-0.005	-0.027 $[-0.041, -0.012]$

Table 16: Historical seed-42 reward-model overoptimization endpoint. Qwen3-4B, locked 200-item GSM8K probe; deltas are endpoint minus reference, and positive IIF denotes degradation.

Mode	**Mitigation**	**IIF**	Δ Proxy	Δ True Acc.	Δ AUC	**KL**
`grpo_rm`	none	$+0.065$	$+1.756$	-0.005	-0.194	0.146
`grpo_rm`	KL anchor	-0.030	$+1.433$	$+0.075$	-0.064	0.085
`grpo_rm`	RM refresh	-0.045	$+2.189$	$+0.090$	-0.025	0.384
`grpo_gt`	none	$+0.005$	-0.118	$+0.025$	-0.011	0.042

for the legacy static-memory endpoint matrices. The updated causal tranche uses a 1,000-item probe and 10,000 bootstrap resamples, with mean 95% dAUC interval width 0.074. Small evaluations are reported with raw counts and coarse intervals.

C.2 Additional Quantitative Diagnostics

Additional diagnostics make failure modes visible without changing the primary estimand. For IIF, useful diagnostics include output length, token distribution, format validity, class balance, calibration, reward-score distribution, solver capability, and downstream error categories across checkpoints. For memory, they include retrieved-neighbor similarity, random-control similarity, semantic-summary length, context truncation, parse failure, and per-condition token cost.

These diagnostics triangulate the execution path. A downstream change that coincides with parser failure or a missing verdict cue is treated as an evaluation-path defect until repaired. A target/witness divergence on common outputs with stable parser rates is stronger evidence of relational incompatibility. Diagnostic plots accompany, but do not replace, the locked primary comparisons.

Secondary audits include a BFCL validation path that fails its coupling gate, a descriptive crossplay comparison with per-cell threshold refitting and no uncertainty interval, and a 120-example qualitative judge audit sampled from 11,000 candidates. These are negative or exploratory diagnostics, not primary IIF estimates. The judge audit categories are exploit, extraction, hard, and harsh. Table 25 summarizes their scope and limitations.

D Resources, Reproducibility and Evidence Ledger

D.1 Resource, Quantization and Reproducibility Registry

D.1.1 Model resources

Table 27 lists the local source/receiver models, reported generators, retrieval encoder, and serving software represented in the result tables, with available resource links and provenance limits.

The Qwen collection and DeepSeek API links above are the only matching official resources listed in the current registry; neither is model-version-specific. Project specifications disagree about Qwen3.8-27B access mode and the DeepSeek-V4-Flash provider, and the available run records do not resolve those conflicts. Mistral and Ministral have version-specific model-card links; their accompanying Mistral paper is family background, not a release citation.

Table 17: Updated three-seed reward-model audit, with seed-level AUC changes and aggregate proxy, task-accuracy, and KL movement.

Condition	ΔAUC by seed	Mean ΔAUC	Mean Δ proxy	Mean Δ true acc.	Mean KL
RM, none	$-0.194/-0.062/-0.036$	-0.097 ± 0.084	$+2.070 \pm 0.949$	$+0.043 \pm 0.042$	0.114 ± 0.031
RM, KL anchor	$-0.064/-0.029/-0.066$	-0.053 ± 0.021	$+3.384 \pm 3.084$	–	–
RM, refresh	$-0.025/+0.041/-0.039$	-0.007 ± 0.042	–	–	–
GT, none	$-0.011/-0.042/-0.038$	-0.030 ± 0.017	$+0.582 \pm 0.613$	$+0.033 \pm 0.010$	0.024 ± 0.016
GT, KL anchor	$-0.020/-0.012/-0.099$	-0.044 ± 0.048	$+0.510 \pm 0.195$	–	–

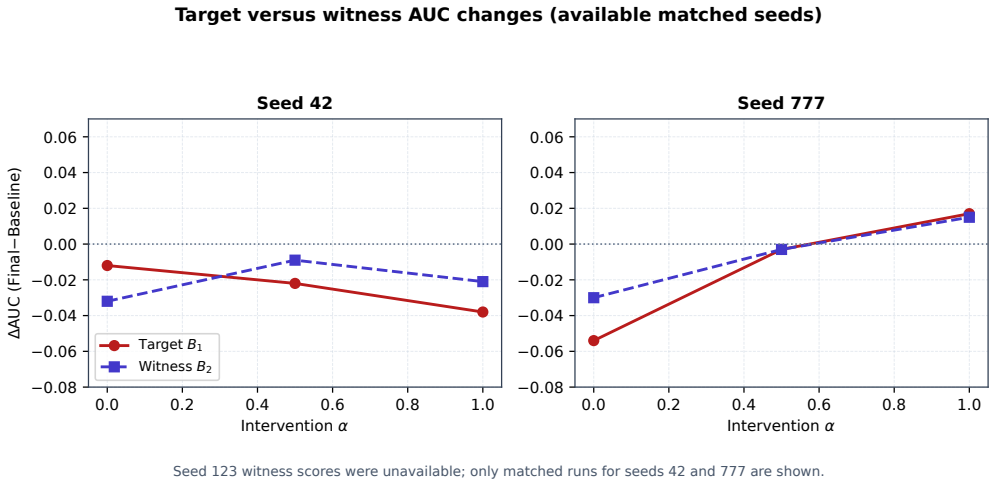

Figure 4: Target judge (B_1) and witness judge (B_2) AUC changes for the available matched runs at intervention strengths $\alpha \in \{0.00, 0.50, 1.00\}$. Seed 123 witness scores were unavailable, so only seeds 42 and 777 are shown; the figure does not represent a complete three-seed witness panel.

D.1.2 Dataset resources

Dataset sizes, evaluation partitions, and construction pools are consolidated in Table 7. The locked 200-item GSM8K probe is a separate IIF audit sample; the MedMCQA development sample is stratified; AIME 2025 contains 30 contemporary items.

D.1.3 Memory resources and retrieval conditions

Memory construction uses Qwen3-Embedding-0.6B Zhang et al. (2025), normalized cosine similarity, top-$k = 5$ retrieval, and a 512-token context budget. The seven conditions are zero-shot, label-only, episodic critique, semantic rubric, hybrid, random retrieval, and raw trajectory. Label-only memory is answer-visible; episodic critique is answer-hidden; random retrieval is the query-irrelevant context control. Construction and evaluation partitions are disjoint, and memory snapshots are versioned separately from model checkpoints.

D.1.4 Training methods and intervention resources

The transfer table compares SFT, GRPO, and a study-defined rubric condition. SFT is contextualized by Ouyang et al. (2022), and GRPO by DeepSeekMath Shao et al. (2024). The reward-model probe uses a Bradley–Terry preference model and frozen judge; proxy-overoptimization motivation is documented by Gao et al. (2023). Ground-truth reward, KL anchoring, reward-model refresh, receiving agent update, and no-mitigation arms are internal controls.

D.1.5 Quantization and precision registry

The active comparisons use an uncompressed BF16 tier and a targeted AWQ INT4 tier Lin et al. (2024). The experiment specification lists the BF16 target hardware as NVIDIA H100 SXM5 and

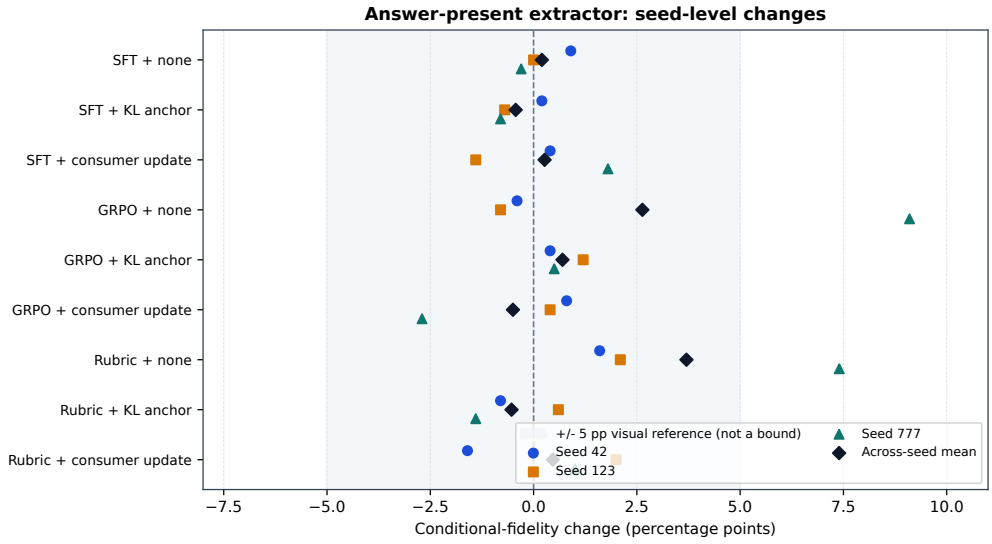

Figure 5: Descriptive seed-level conditional-fidelity changes for the answer-present extractor across nine update conditions. Points show individual seeds and diamonds show across-seed means. The shaded ± 5-percentage-point band is a visual reference, not an equivalence bound; condition-level interval estimates are not reported.

Table 18: Historical seed-42 solver–extractor drift endpoints. Fidelity is evaluated only when the reference solver is correct at both endpoints.

Training	Mitigation	IIF Fidelity	IIF Pipeline	KL Divergence
SFT	none	$+0.009$	-0.040	0.114
SFT	Receiving agent update	$+0.004$	-0.020	0.140
SFT	KL anchor	$+0.002$	-0.065	0.107
GRPO	none	-0.004	-0.065	0.036
GRPO	KL anchor	$+0.004$	-0.030	0.024
GRPO	Receiving agent update	$+0.008$	-0.055	0.033
Rubric	none	$+0.016$	-0.040	0.247
Rubric	KL anchor	-0.008	-0.020	0.017
Rubric	Receiving agent update	-0.016	-0.085	0.111

NVIDIA A100 SXM4 (80 GB each), and the AWQ INT4 targets as NVIDIA GeForce RTX 4090 (24 GB) and NVIDIA RTX A6000 (48 GB). Active run records do not identify per-run device assignments, so these are documented targets rather than confirmed execution provenance. Transfer comparisons remain within precision. Locally measured baselines are reported in Table 28; Figure 12 summarizes approximate AWQ offsets of 1.3–2.0 points on GSM8K, MATH-500, GPQA, and MedMCQA, and 3.3–3.4 points on AIME 2025.

D.1.6 Execution and artifact registry

Local and gateway comparisons use $T = 0.0$, top-$p = 1.0$, top-$k = -1$, repetition penalty 1.0, and seed 42 unless a table records a provider-specific setting. The execution registry schema requires model identifiers and revisions, provider, prompt version, dataset revision, split manifest, normalization rule, memory snapshot, parser, scorer, environment lock, hashes, completion counts, failures, configuration hashes and artifact paths. These fields should be described as present only when the corresponding run manifest is supplied.

Table 19: Three-seed solver–extractor fidelity audit. Values are means and standard deviations across seeds 42, 123, and 777. Paired denominators count examples with a correct solver answer at both the reference and step-200 endpoints; seed order is 42/123/777.

Condition	Seed values	Mean ± SD	n_{pair} (42/123/777)
SFT + none	$+0.009/+0.000/-0.003$	$+0.002 \pm 0.006$	145/141/145
SFT + KL anchor	$+0.002/-0.007/-0.008$	-0.004 ± 0.005	146/144/140
SFT + receiving-agent update	$+0.004/-0.014/+0.018$	$+0.003 \pm 0.016$	141/147/142
GRPO + none	$-0.004/-0.008/+0.091$	$+0.026 \pm 0.056$	145/152/144
GRPO + KL anchor	$+0.004/+0.012/+0.005$	$+0.007 \pm 0.004$	147/150/145
GRPO + receiving-agent update	$+0.008/+0.004/-0.027$	-0.005 ± 0.019	148/148/145
Rubric + none	$+0.016/+0.021/+0.074$	$+0.037 \pm 0.032$	143/146/146
Rubric + KL anchor	$-0.008/+0.006/-0.014$	-0.005 ± 0.010	149/134/145
Rubric + receiving-agent update	$-0.016/+0.020/+0.010$	$+0.005 \pm 0.019$	149/150/148

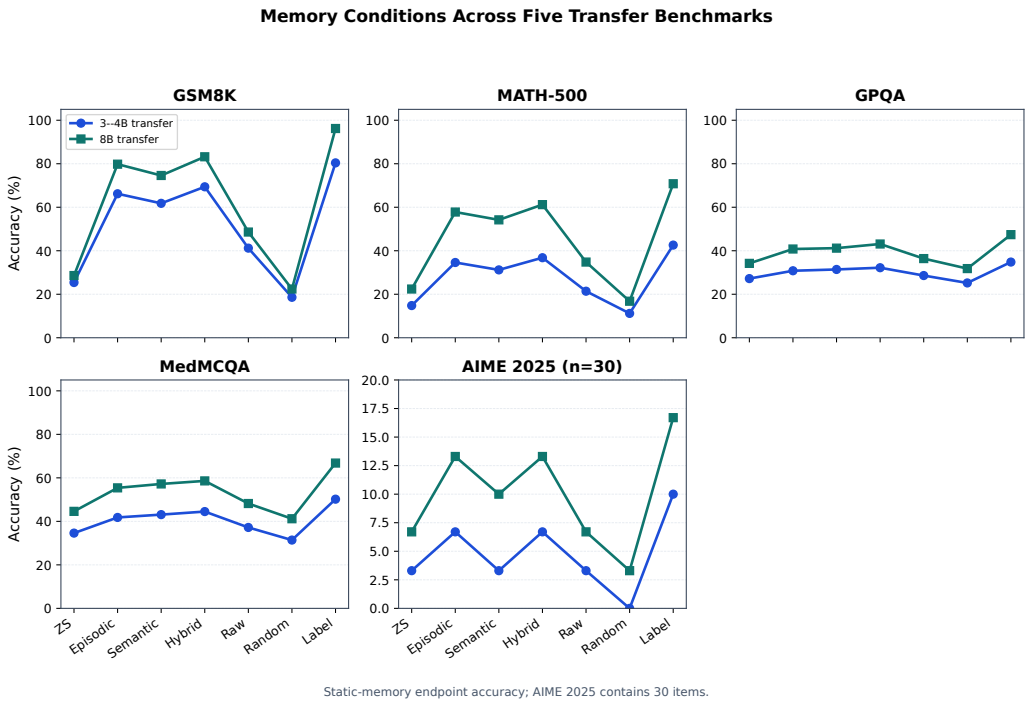

Figure 6: Static-memory conditions across four transfer benchmarks. The curves compare the 3–4B and 8B counterparty directions across seven conditions, with receiving agent endpoint accuracy shown for each memory condition.

D.2 Reproducibility and Artifact Documentation

For release, each reported number should map to a machine-readable artifact containing dataset revision, split hash, source and receiving-agent models, critic or judge, prompt version, seed, decoding, precision, memory snapshot, retrieval settings, parser, scorer and content hashes. The release package should separate source code, environment specification, prompts, data-access instructions, manifests, memory snapshots where permitted, per-example predictions and table-generation scripts. Table 29 maps the reported result families to evidence recoverable in the current workspace and identifies gaps; this mapping does not establish that those files are included in a release package. API logs should exclude credentials while retaining model identifier, date, token counts, retries and parsing outcome.

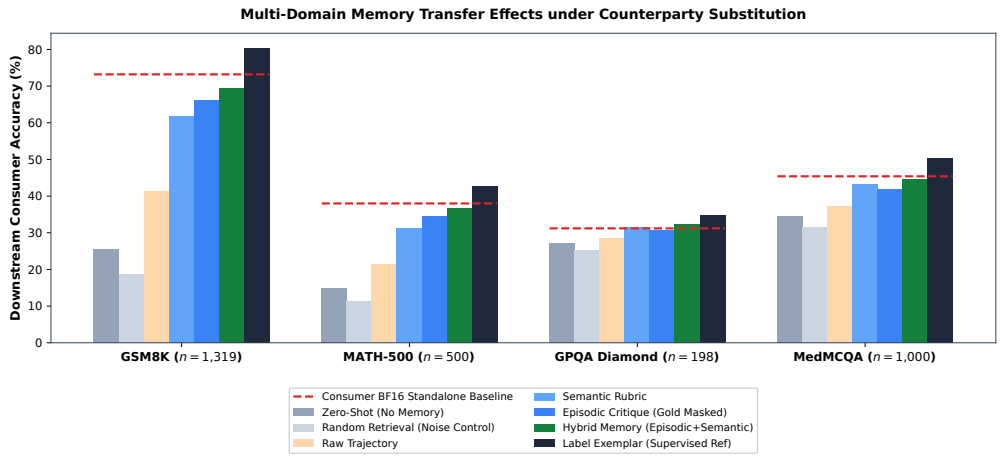

Figure 7: Static-memory endpoint accuracy across four benchmarks. Hybrid and answer-visible label-exemplar conditions are shown separately.

Table 20: Legacy, provisional seven-configuration static-memory endpoint accuracy across the listed partner substitutions. Memory IDs: 0 = zero-shot; 1 = answer-hidden episodic critique; 2 = semantic rubric; 3 = hybrid; 4 = raw trajectory; 5 = random retrieval; 6 = answer-visible label exemplar. These endpoint values do not estimate memory mitigation under drift.

Transfer pairing	Memory ID	GSM8K	GPQA	MATH-500	MedMCQA	AIME 2025
Qwen3-4B → Llama3.2-3B	0	25.4	27.2	14.8	34.6	3.3
Qwen3-4B → Llama3.2-3B	1	66.2	30.8	34.6	41.8	6.7
Qwen3-4B → Llama3.2-3B	2	61.8	31.4	31.2	43.1	3.3
Qwen3-4B → Llama3.2-3B	3	69.4	32.2	36.8	44.5	6.7
Qwen3-4B → Llama3.2-3B	4	41.2	28.6	21.4	37.2	3.3
Qwen3-4B → Llama3.2-3B	5	18.6	25.2	11.2	31.4	0.0
Qwen3-4B → Llama3.2-3B	6	80.4	34.8	42.6	50.2	10.0
Llama3.2-3B → Qwen3-4B	0	24.1	26.8	13.9	33.8	3.3
Llama3.2-3B → Qwen3-4B	1	65.4	30.2	33.8	41.2	6.7
Llama3.2-3B → Qwen3-4B	2	60.9	30.8	30.6	42.4	3.3
Llama3.2-3B → Qwen3-4B	3	68.6	31.6	35.9	43.8	6.7
Llama3.2-3B → Qwen3-4B	4	40.5	28.1	20.8	36.5	3.3
Llama3.2-3B → Qwen3-4B	5	17.9	24.8	10.6	30.8	0.0
Llama3.2-3B → Qwen3-4B	6	79.2	34.1	41.8	49.4	10.0
Qwen3-8B → Llama3.1-8B	0	28.6	34.2	22.4	44.6	6.7
Qwen3-8B → Llama3.1-8B	1	79.8	40.8	57.8	55.4	13.3
Qwen3-8B → Llama3.1-8B	2	74.6	41.2	54.2	57.2	10.0
Qwen3-8B → Llama3.1-8B	3	83.2	43.1	61.2	58.6	13.3
Qwen3-8B → Llama3.1-8B	4	48.6	36.4	34.8	48.2	6.7
Qwen3-8B → Llama3.1-8B	5	22.4	31.8	16.8	41.2	3.3
Qwen3-8B → Llama3.1-8B	6	96.2	47.4	70.8	66.8	16.7
Llama3.1-8B → Qwen3-8B	0	27.8	33.5	21.6	43.8	6.7
Llama3.1-8B → Qwen3-8B	1	78.9	40.2	56.9	54.6	13.3
Llama3.1-8B → Qwen3-8B	2	73.8	40.6	53.4	56.4	10.0
Llama3.1-8B → Qwen3-8B	3	82.4	42.4	60.4	57.8	13.3
Llama3.1-8B → Qwen3-8B	4	47.9	35.8	34.1	47.5	6.7
Llama3.1-8B → Qwen3-8B	5	21.8	31.2	16.2	40.5	3.3
Llama3.1-8B → Qwen3-8B	6	95.4	46.8	69.9	65.9	16.7

D.3 Evidence and Claim Ledger

The crosswalk distinguishes run-level analysis records from manuscript summaries and unavailable provenance. Interpretation should remain tied to the measured outcome and comparison design.

This crosswalk records what can be traced from the present workspace without promoting summary-only records into independently verified empirical evidence. The static-memory and generated-trace provenance gaps remain open.

Table 21: Seven-condition direct-prediction results under a standardized five-shot CoT harness.

Benchmark	Receiving agent model	n	Metric	CoT base	Label	Episodic	Semantic	Hybrid	Random	Raw
GPQA Diamond	Qwen3-4B (BF16 local)	198	Accuracy	48.5	50.7	53.4	51.6	55.1	47.8	49.6
GPQA Diamond	Llama3.2-3B (BF16 local)	198	Accuracy	45.2	46.5	49.0	47.3	50.2	44.4	46.0
MATH-500	Qwen3-4B (BF16 local)	500	Accuracy	44.6	46.1	48.7	47.2	50.4	43.7	45.5

Table 22: Cross-model partner-substitution endpoint accuracy by training method, model tier, benchmark, and precision.

Partner swap ($A \to A'$)	Tech.	Precision	GSM8K	GPQA	MATH-500	MedMCQA	AIME 2025
3–4B parameter class (n = 1,319 / 198 / 500 / 1,000 / 30)							
Qwen3-4B → Llama3.2-3B	SFT	BF16	25.4	27.2	14.8	34.6	3.3
Qwen3-4B → Llama3.2-3B	GRPO	BF16	46.8	29.5	25.4	40.2	6.7
Qwen3-4B → Llama3.2-3B	Rubric	BF16	13.2	22.4	8.6	25.1	0.0
Llama3.2-3B → Qwen3-4B	SFT	BF16	24.1	26.8	13.9	33.8	3.3
Llama3.2-3B → Qwen3-4B	GRPO	BF16	45.2	28.8	24.2	39.4	3.3
Llama3.2-3B → Qwen3-4B	Rubric	BF16	12.5	21.8	7.8	24.2	0.0
Qwen3-4B → Llama3.2-3B	SFT	AWQ INT4	23.8	25.4	13.2	32.7	0.0
Qwen3-4B → Llama3.2-3B	GRPO	AWQ INT4	44.1	27.6	23.8	38.5	3.3
8B parameter class (n = 1,319 / 198 / 500 / 1,000 / 30)							
Qwen3-8B → Llama3.1-8B	SFT	BF16	28.6	34.2	22.4	44.6	6.7
Qwen3-8B → Llama3.1-8B	GRPO	BF16	55.4	38.6	41.2	52.8	10.0
Qwen3-8B → Llama3.1-8B	Rubric	BF16	15.8	28.4	12.6	33.2	3.3
Llama3.1-8B → Qwen3-8B	SFT	BF16	27.8	33.5	21.6	43.8	6.7
Llama3.1-8B → Qwen3-8B	GRPO	BF16	54.1	37.8	40.4	51.9	6.7
Llama3.1-8B → Qwen3-8B	Rubric	BF16	14.9	27.6	11.8	32.4	3.3
Qwen3-8B → Llama3.1-8B	SFT	AWQ INT4	26.9	32.8	21.0	42.9	3.3
Qwen3-8B → Llama3.1-8B	GRPO	AWQ INT4	53.2	36.9	39.6	50.8	6.7

E Limitations, Ethics and Protocol-Only Extensions

E.1 Direct Test of Memory Under Upstream Drift

Protocol status: design specification only; no completed outcome values are reported. A future memory-under-drift test would hold upstream trajectories and evaluation examples fixed across memory conditions. The planned arms are no memory, static reference memory, refreshed memory constructed from the changed partner, swapped-partner memory, random memory and where feasible, a weight-based adaptation control. Own, swapped, pooled and generic memory banks would be matched on record count, token budget, retrieval budget and construction constraints so that source identity is not confounded with context volume.

For memory state M_m, let $P_m(t)$ be the frozen receiving agent's task validity on the cached upstream outputs at partner state t. The degradation under that memory condition is

$$\Delta_{\mathrm{IIF},m} = P_m(0) - P_m(t). \tag{10}$$

The mitigation contrast is

$$\mathrm{Mitigation}_m = \Delta_{\mathrm{IIF},\text{no memory}} - \Delta_{\mathrm{IIF},m}, \tag{11}$$

so positive values would indicate less degradation than the no-memory condition. A static-memory endpoint gain alone does not estimate this contrast because it does not compare the same changed-partner outputs across memory states. Any future implementation should use paired uncertainty and separate construction from evaluation data.

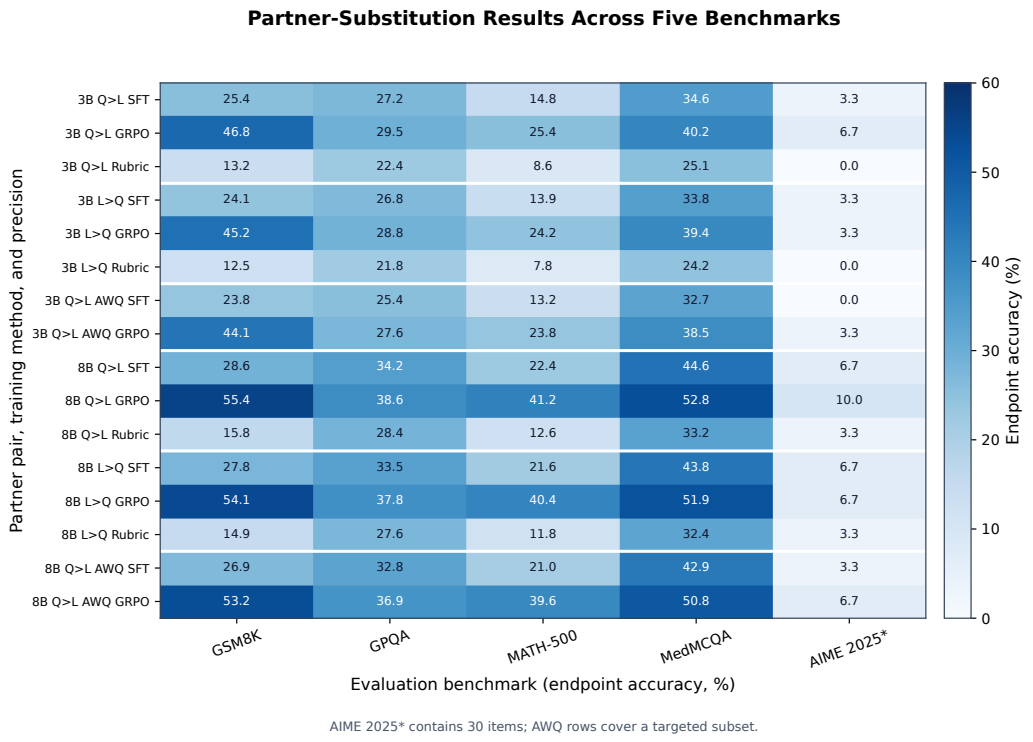

Figure 8: Partner-substitution endpoint accuracy by training method, model tier, and benchmark.

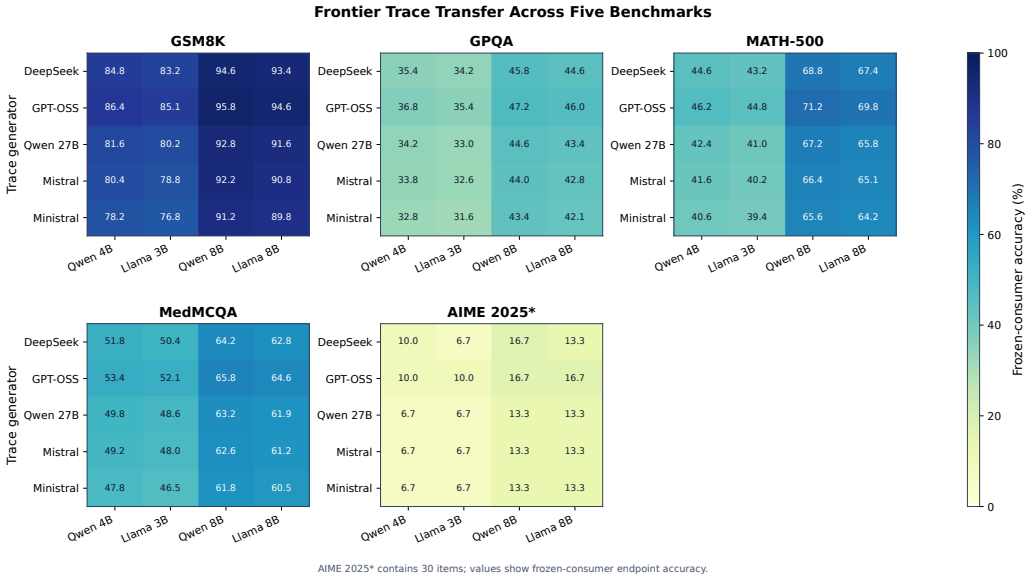

Figure 9: Frontier/API trace transfer to frozen receiving agents on GSM8K and MATH-500. Cells show final receiving agent endpoint accuracy.

E.2 Memory Ablations, Sensitivity and Cost

Protocol status: memory ablations and cost analyses remain unrun; no outcome values are reported. The planned dimensions are episodic versus semantic content, top-k, bank size, summary length, construction seed, retrieval key, context order and prompt budget. Hyperparameters would be se-

Table 23: Frontier/API trace transfer to frozen receiving agents. Values are receiving agent endpoint accuracies.

Upstream generator	Frozen receiving agent	GSM8K	GPQA	MATH-500	MedMCQA	AIME 2025
DeepSeek-V4-Flash generator						
DeepSeek-V4-Flash	Qwen3-4B	84.8	35.4	44.6	51.8	10.0
DeepSeek-V4-Flash	Llama3.2-3B	83.2	34.2	43.2	50.4	6.7
DeepSeek-V4-Flash	Qwen3-8B	94.6	45.8	68.8	64.2	16.7
DeepSeek-V4-Flash	Llama3.1-8B	93.4	44.6	67.4	62.8	13.3
GPT-OSS-120B generator						
GPT-OSS-120B	Qwen3-4B	86.4	36.8	46.2	53.4	10.0
GPT-OSS-120B	Llama3.2-3B	85.1	35.4	44.8	52.1	10.0
GPT-OSS-120B	Qwen3-8B	95.8	47.2	71.2	65.8	16.7
GPT-OSS-120B	Llama3.1-8B	94.6	46.0	69.8	64.6	16.7
Qwen3.8-27B generator						
Qwen3.8-27B	Qwen3-4B	81.6	34.2	42.4	49.8	6.7
Qwen3.8-27B	Llama3.2-3B	80.2	33.0	41.0	48.6	6.7
Qwen3.8-27B	Qwen3-8B	92.8	44.6	67.2	63.2	13.3
Qwen3.8-27B	Llama3.1-8B	91.6	43.4	65.8	61.9	13.3
Mistral-Small-24B-Instruct-2501 generator						
Mistral-Small-24B-Instruct-2501	Qwen3-4B	80.4	33.8	41.6	49.2	6.7
Mistral-Small-24B-Instruct-2501	Llama3.2-3B	78.8	32.6	40.2	48.0	6.7
Mistral-Small-24B-Instruct-2501	Qwen3-8B	92.2	44.0	66.4	62.6	13.3
Mistral-Small-24B-Instruct-2501	Llama3.1-8B	90.8	42.8	65.1	61.2	13.3
Ministral-3-14B-Instruct-2512 generator						
Ministral-3-14B-Instruct-2512	Qwen3-4B	78.2	32.8	40.6	47.8	6.7
Ministral-3-14B-Instruct-2512	Llama3.2-3B	76.8	31.6	39.4	46.5	6.7
Ministral-3-14B-Instruct-2512	Qwen3-8B	91.2	43.4	65.6	61.8	13.3
Ministral-3-14B-Instruct-2512	Llama3.1-8B	89.8	42.1	64.2	60.5	13.3

Table 24: Standalone accuracy reference summaries for named generator models under deterministic greedy decoding.

Model	Deployment	GSM8K	GPQA	MATH-500
Qwen3.8-27B	BF16 local	94.2	64.9	78.4
DeepSeek-V4-Flash	API-hosted	95.8	69.7	88.2
GPT-OSS-120B (high reasoning)	API-hosted	98.2	78.6	94.5
Mistral-Small-24B-Instruct-2501	API-hosted	91.4	58.2	71.6

lected on validation data before test evaluation. Planned cost accounting includes critic calls, embeddings, storage, retrieval latency, prompt tokens, receiving-agent latency, API usage and refresh labels.

E.3 Extended Limitations and Boundary Conditions

The framework applies to systems where an upstream message distribution and downstream receiving agent can be isolated with frozen probes. Feedback loops, hidden tool state, human intervention and nonstationary environments complicate the decomposition. The historical endpoint probes use a 200-item GSM8K panel, while the updated causal tranche uses three seeds and 1,000 items with imperfect forward-KL matching and no monotone dose-response. AWQ coverage is targeted and AIME 2025 contains 30 items. External memory introduces stale records, contradiction, contamination, answer leakage, retrieval misses, prompt pressure and poisoning risks.

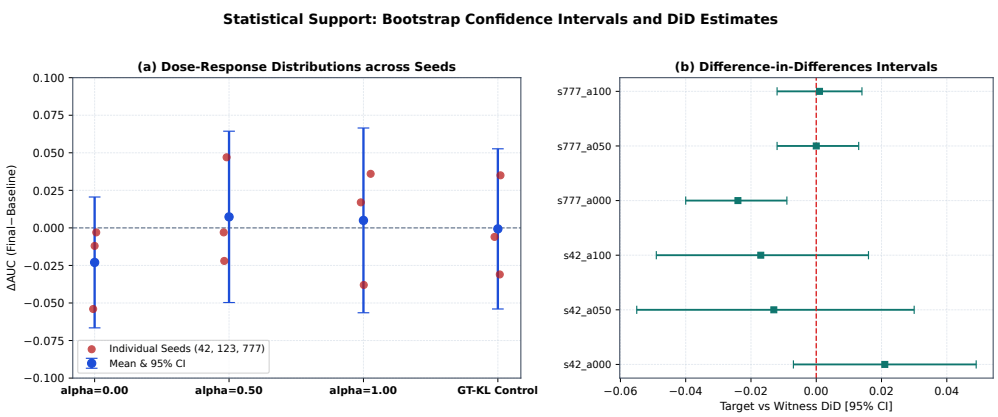

Figure 10: (a) Bootstrap distributions of ΔAUC across four intervention levels for seeds 42, 123, and 777. No monotone dose-response is visible: the mean estimates include zero for all conditions, and individual seed trajectories diverge in direction. (b) Target-versus-witness difference-in-differences intervals for six seed-$\times$-α cells, none of which show a significant positive DiD, confirming the bounded-null finding.

Table 25: Negative and exploratory verifier, crossplay, judge, and memory measurements. These rows are not primary claim-bearing results.

Measurement	Scale	Observed value	Comparison detail
Natural-drift verifier	Historical trial	Edge $+0.000$, AUC 0.500	Approval exceeded 99.5%; matched always-approve baseline
Reward-model judge	Seeds 42/777	Edge $+0.040/+0.065$; AUC $0.822/0.849$	Both exceeded the constant-baseline gate
BFCL validation	$n = 133$	AUC 0.4605 vs. baseline 0.5038	Validation AUC below baseline
Crossplay matrix	60 cells, $n = 1000$	Own AUC 0.8139 vs. other 0.7988	Threshold refit per cell
T11 judge audit	120 of 11,000	Exploit 21; extraction 50; hard 30; harsh 19	Counts by judge-output category
GPQA memory results	$n = 10$ per condition	0–2 correct in Qwen conditions	Accuracy counts by condition

E.4 Broader Impacts, Ethics and Risk

Reliable inter-agent interfaces may improve oversight, tool-use safety and system maintenance. Identifying weak verifier features can also make it easier to exploit a learned judge, so the paper does not release attack prompts or deployment-specific evasion recipes. Releases must respect dataset licenses, provider terms, model licenses and restrictions on generated traces. Compute reporting should include local accelerator time, API usage, storage and approximate energy accounting.

F Extended Related-Work Survey

F.1 Extended Related-Work Survey

This appendix records the broader paper-to-gap comparison behind Section 2. The organizing question is what each work trains or evaluates and whether its intervention measures the reliability of a fixed receiving agent when its upstream partner changes. The table retains representative papers with local PDFs and catalog entries; it is a survey aid, not evidence for the empirical claims in this paper.

F.1.1 Full original literature taxonomy

The following taxonomy preserves the original broad survey categories and comparison fields. It is retained in the Appendix to keep the main Related Work focused on the closest novelty neighbors while preserving the wider literature map.

Table 26: Completed memory-transfer scope and protocol-only comparison dimensions. Rows marked protocol-only have no reported outcome values.

Comparison	Conditions and controls
Memory benchmark evaluation (protocol only)	Corrected GSM8K, GPQA, MATH-500 and AIME runs with fixed parser, truncation, validity gate and paired outputs; no corrected outcomes reported
Memory under drift (protocol only)	No memory, weight-only, static, refreshed and weight-plus-memory on identical cached trajectories; no outcomes reported
Partner-specific memory (protocol only)	Own, swapped, pooled and generic banks matched for size and token budget; no outcomes reported
Validation and seed hierarchy (protocol only)	Validation-based epoch selection with independent construction/evaluation seeds; no outcomes reported
Refresh and staleness (protocol only)	Drift age, refresh interval, recovery and maintenance cost; no outcomes reported
Memory ablations and costs (protocol only)	Retrieval key, top-k, bank size, context budget, equal-token controls, latency and API cost; no outcomes reported

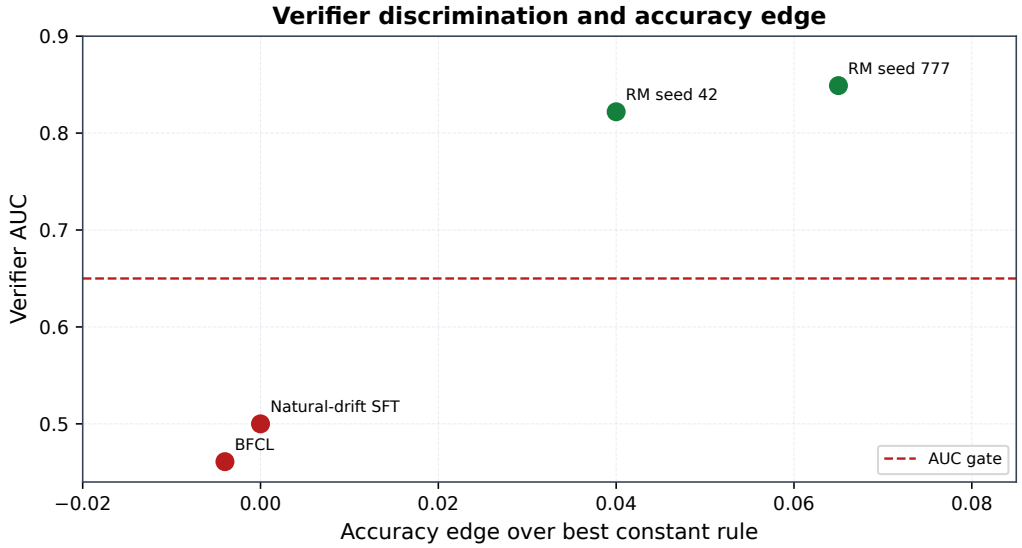

Figure 11: Coupling-validity measurements. Early natural-drift and BFCL verifier AUCs are shown against the constant-baseline criterion; reward-model judge AUCs exceed that criterion.

F.1.2 Co-training, verification and optimization

Solver–verifier research separates proposal from assessment and learned-evaluator work studies the quality of those assessments. Training verifiers on mathematical solutions, process-supervision work and Prometheus-style fine-grained evaluation establish useful roles and benchmarks for learned judges Cobbe et al. (2021); Lightman et al. (2024); Kim et al. (2024). Their focus is verifier or evaluation performance, rather than a longitudinal intervention on the producer while holding the receiving agent fixed.

Co-training papers study collaborative adaptation directly. MAPoRL uses post-training for collaborative language models; WaltzRL jointly trains conversation and feedback agents for safety; CoMAS optimizes interaction rewards; and Dr. MAS targets stability in multi-agent reinforcement learning Park et al. (2025); Zhang et al. (2026b); Xue et al. (2026); Feng et al. (2026a). These works motivate controls for co-adaptation and policy movement, but their primary estimands concern collaboration, safety, or training stability rather than task-validity loss in a frozen receiving agent after partner replacement.

F.1.3 Memory and interference

MemoryAgentBench evaluates incremental multi-turn memory capabilities, whereas CoMAM studies joint optimization of extraction, profile and retrieval components Hu et al. (2026b); Mao et al. (2026). Their evaluation units are memory competencies and memory-system optimization; nei-

Table 27: Model and software resources represented in the reported result tables.

Resource	Role	Provenance
Qwen3-4B Yang et al. (2025a)	Local source/receiver	Qwen3 paper and official model card
Llama 3.2 3B Instruct Grattafiori et al. (2024)	Local source/receiver	Llama 3 paper and official model card
Qwen3-8B Yang et al. (2025a)	Local source/receiver	Qwen3 paper and official model card
Llama 3.1 8B Instruct Grattafiori et al. (2024)	Local source/receiver	Llama 3 paper and official model card
Qwen3.8-27B	Generator/critic; deployment unresolved	Official Qwen model collection; no model-specific card or run revision is listed
DeepSeek-V4-Flash	Generator; provider unresolved	Official DeepSeek API documentation; no row-aligned version or request record is available
GPT-OSS-120B	Generator; deployment unresolved	Official model card; no row-aligned gateway record is available
Mistral-Small-24B-Instruct-2501 Jiang et al. (2023)	Generator; deployment unresolved	General Mistral background paper and version-specific official model card
Ministral-3-14B-Instruct-2512 Jiang et al. (2023)	Generator; deployment unresolved	General Mistral background paper and version-specific official model card
Qwen3-Embedding-0.6B Zhang et al. (2025)	Retrieval encoder	Embedding paper and official model card
vLLM 0.6.3 Kwon et al. (2023)	Local inference engine	PagedAttention paper and official release

Table 28: Locally measured baseline references and quantization calibration. These values anchor transfer and memory calculations to the matching receiving agent and precision.

Model	Precision	GSM8K	MATH-500	GPQA	MedMCQA	AIME 2025
Qwen3-4B	BF16	74.6%	39.2%	31.8%	46.2%	6.7%
Qwen3-4B	AWQ INT4	72.8%	37.4%	29.8%	44.5%	3.3%
Llama3.2-3B	BF16	73.2%	38.0%	31.2%	45.4%	6.7%
Llama3.2-3B	AWQ INT4	71.5%	36.2%	29.2%	43.6%	3.3%
Qwen3-8B	BF16	89.4%	64.6%	42.4%	60.4%	13.3%
Qwen3-8B	AWQ INT4	88.1%	62.8%	40.9%	58.7%	10.0%
Llama3.1-8B	BF16	88.2%	63.5%	41.6%	59.6%	13.3%
Llama3.1-8B	AWQ INT4	86.8%	61.7%	40.1%	57.9%	10.0%
Hosted-model reference	API-hosted	96.5%	92.4%	72.5%	71.8%	40.0%

ther is the matched comparison of no memory, static memory and refreshed memory on identical changed-partner trajectories required to establish causal IIF mitigation.

ICF-Bench studies within-model interference from conflicting in-context information and cognitive work reviews interference and inhibition mechanisms Qian et al. (2026); Kliegl & Bäuml (2021). These works ground the broader vocabulary of interference and forgetting, but the intervention is not a sender-receiver boundary with an upstream partner change and a fixed downstream component.

F.1.4 Drift, communication and system evaluation

Agent Drift examines behavioral degradation over extended interactions; MasDrift and Constraint Drift focus on authorization or safety-constraint preservation across multi-agent trajectories Rath (2026); Xu et al. (2026b); Li et al. (2026). ChannelGuard studies unsafe composition, while the latent-communication audit tests whether relayed representations causally communicate Hossain et al. (2026); Cheng et al. (2026). Simulator-collapse work further shows that multi-agent behavior can depend on the learned or frozen environment in which policies interact Yu et al. (2026b). These lines address important forms of drift, safety, communication and evaluation validity. Their outcomes and interventions differ from the task-grounded, validity-gated reliability of an unchanged receiving agent under a controlled change to its upstream source.

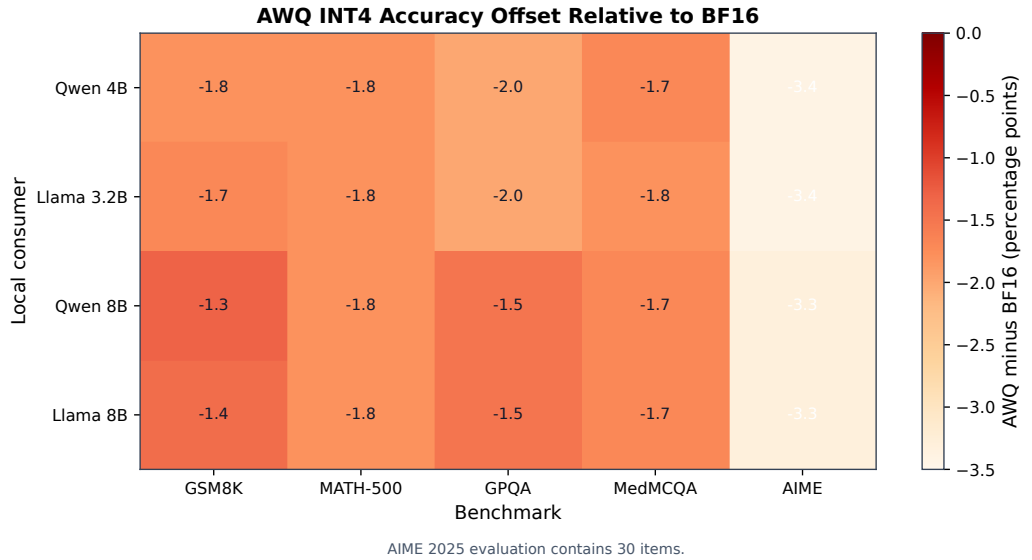

Figure 12: BF16-to-AWQ INT4 accuracy offsets for the four local receiving agents and five benchmarks. The AIME offset is largest; its evaluation contains 30 items.

Table 29: Result-to-evidence crosswalk for the active manuscript. Evidence locations refer to the current workspace; availability here does not establish inclusion in the submitted or released artifact package.

Result family	Reported result	Evidence recoverable in workspace	Boundary / missing provenance
Causal tranche (T1/T2/T3/T5)	1,000-item judge AUC, KL-matched contrasts, target-witness DiD and dose response	`analyze_causal.py` and its `CAUSAL_ANALYSIS.md` run-level report; T1 coverage and intervals are reproduced in Table 14	Run-level report and analysis code are present; raw probe scores and inclusion in a release package are not established here. T3 witness coverage is limited to seeds 42 and 777, plus a separate seed-777 GT-KL control.
Reward-model audit	Three-seed RM-versus-GT judge-AUC comparison	Seed-level values and paired comparison are shown in Table 17; the paired t-test is across seeds	This table is the mapped evidence; a machine-readable run manifest and API/request logs are not identified in the active record.
Static-memory transfer	Seven-condition endpoint accuracy under partner substitution	Legacy endpoint summaries are reported in Table 20 and the associated figure	Provisional summaries only: row-aligned run manifests, memory snapshots and per-example outputs are not mapped here. These comparisons do not estimate IIF or memory mitigation under drift.
Generated-trace results	Per-generator endpoint means across local receiving agents	The generator-by-receiver summary is reported in Table 23	Available GPQA manifests correspond to different runs; row-aligned request records do not resolve provider, deployment, version, or snapshot for the reported table. No provider-specific attribution is made.
Solver–extractor audit	Three-seed conditional-fidelity changes and paired denominators	Seed-level changes and denominators are shown in Table 19	Descriptive values only; no condition-level intervals or significance tests are reported. The table does not establish artifact-package completeness.

Table 30: Information recorded for each reported result.

Record field	Reported information
Metric and estimate	Outcome, units, sample size and uncertainty
Protocol	Model, condition, prompt, decoding, precision and dataset split
Validity checks	Baseline comparison, parser behavior and execution-path controls
Source artifact	Result records and table-generation inputs
Interpretation	Comparison-specific meaning and scope

Table 31: What the completed memory setup does and does not establish about IIF.

Evidence	Comparison performed	Measured outcome	IIF interpretation
Static memory transfer (completed; provisional)	Seven memory conditions under partner substitution; 4 directions and 5 benchmark panels	Receiving-agent endpoint accuracy; relative endpoint accuracy summarized over 8 GSM8K/MATH-500 cells	Does not estimate IIF: no matched upstream drift or paired baseline-to-drift change
Memory under identical drift (planned protocol M5)	No-memory, static, refreshed and weight-plus-memory arms on identical trajectories	No outcomes reported; protocol rows are unrun	IIF and memory mitigation are not established for this setup

Table 32: Survey taxonomy of research lines relevant to interference-induced forgetting. Each row identifies the principal technique family, the interacting roles, the mechanism that couples them, the usual evaluation target and the limitation that motivates a validity-gated interface analysis. Paper references use bibliography citations.

Line	Examples	Roles	Mechanism	Evaluation	IIF relevance / limit
Coordination frameworks	AutoGen Wu et al. (2023); CAMEL Li et al. (2023); MetaGPT Hong et al. (2024)	Planner, solver, critic	Role prompts, dialogue, tools	Coding, QA, workflows	Modular teams; fixed-roster validity usually untested.
Long-context collaboration	Chain of Agents Zhang et al. (2024); Geometry Furuya & Kitagawa (2025); KVComm	Relays, sequential agents	Long-context/KV handoff, graph dialogue	Long-context synergy	Receiver calibration under sender change is untested.
Solver–verifier decomposition	Li et al. (2024a); Sun et al. (2026); Lu et al. (2025); Cobbe et al. (2021)	Solver and verifier	Candidate scores, labels, process checks	Math and verifier accuracy	Basis for frozen-receiving agent tests; no partner-change estimand.
Verifier-guided search	Lai et al. (2026); Yang et al. (2025b); Lightman et al. (2024)	Solver, critic, verifier	Critiques, rewards, search	Math and test-time scaling	Judge optimization can damage validity.
Learned judges	Laddha et al. (2026); Kim et al. (2024); Zheng et al. (2023)	Generator and evaluator	Rubrics, scalar/pairwise scores	Agreement and model comparison	Calibration bias; producer-shift testing is missing.
Debate and deliberation	Hu et al. (2026a); Wu et al. (2026); Yu et al. (2026a); Bąba & Chudziak (2026); Niu & Zhang (2026); Pecher (2026); Feng et al. (2026b); Nguyen et al. (2026); Du et al. (2023)	Solvers, debaters, moderator	Arguments, voting, disagreement	Factuality, math, social reasoning	Aggregate success often replaces interface analysis.
Co-training and post-training	Dong et al. (2024); Zheng et al. (2024); Hong et al. (2025); Estornell et al. (2025); Park et al. (2025); Pulici et al. (2026); Zhang et al. (2026b); Xue et al. (2026)	Partner policies, critic	Shared trajectories, RL, adapters	Collaboration and policy improvement	Frozen-partner validity is rarely isolated.
Multi-agent reinforcement learning	Yu et al. (2026b); Li et al. (2024b); Sela (2026); Feng et al. (2026a)	Policies, simulators, communicators	Shared rewards, messages, feedback	Coordination and stability	Co-adaptation can collapse; our receiving agent is frozen.
Reward overoptimization	Shao et al. (2024); Bai et al. (2022); Gao et al. (2023)	Policy, preference model, verifier	Proxy reward, preferences, KL	Gold versus proxy score	Direct basis for target-sensitive judge tests.
Continual and in-context forgetting	Li & Hoiem (2017); Kirkpatrick et al. (2017); Qian et al. (2026); Kliegl & Bäuml (2021)	Updated model or one model processing conflicting context	Parameter retention or in-context interference	Continual learning and selective forgetting	Differs from a frozen receiving agent under an upstream partner change.
External retrieval	Brown et al. (2020); Lewis et al. (2020)	Generator, retriever, receiving agent	Indexes, demonstrations, context	Knowledge QA and transfer	Separates generic assistance from partner recovery.
Episodic agent memory	Ke et al. (2026); Shi et al. (2026); Park et al. (2023); Shinn et al. (2023); Hu et al. (2026b); Zhang et al. (2026a)	Agent, reflection, memory store	Trajectories, critique, retrieval	Lifelong and domain tasks	Leakage, stale records and source identity matter.
Semantic and reflective memory	Zhang et al. (2026c); Jin et al. (2026); Rajib et al. (2026); Mao et al. (2026)	Planner, memory manager, solver	Summaries, rules, weighted records	Planning and collaboration	Compression may lose partner-specific cues.
Agentic and skill memory	Packer et al. (2023); Xu et al. (2026a); Wang et al. (2023)	Agent, memory manager, tools	Skills, linked notes, context management	Lifelong and long-context tasks	Differs from frozen benchmark receiving agents.
Communication robustness	Ye et al. (2026); Cheng et al. (2026); Zhang & Emu (2026); Hossain et al. (2026)	Sender, receiver, safety evaluator	Latent channels, KV, composition	Causal and safety tests	Communication must be established; IIF tests observable validity.
Interface and format contracts	Huang et al. (2025); Ansell & Toney-Wails (2026); Lou et al. (2026); Zhao et al. (2026); Huang et al. (2026)	Solver, parser, tool caller	Schemas, AST calls, tool state	Function-call validity	Format-sensitive receiving agents are a key IIF follow-up.
Heterogeneous teams	Sharif & Rehman (2026); Lou et al. (2026); Niu & Zhang (2026)	Different models and roles	Cross-model messages and routing	Graph, debate, reasoning	Partner substitution is the maintenance stress test.
Underspecification and shortcuts	Wang et al. (2026a); Huang & Deng (2026); D'Amour et al. (2022)	Pipeline, solver, evaluator	Heuristics and output conventions	Robustness under change	Equal scores need not imply equal interface behavior.
Evaluation and failure attribution	Lee et al. (2025); Chen et al. (2026); Wang et al. (2026b); Khatri et al. (2026); Rasker et al. (2026)	Agents, evaluator, auditor	Metrics, calibration, traces	Auditability and diagnosis	Supports gates, effect definitions and provenance.

Table 33: Selected papers, their evaluated objects and their relation to the frozen-receiving agent interface gap.

Technique / papers	Agent roles	Communication / memory mechanism	Evaluation target	Remaining gap
Agent orchestration: AutoGen, CAMEL, MetaGPT Wu et al. (2023); Li et al. (2023); Hong et al. (2024)	Planner, worker, critic, and executor	Role prompts, dialogue, and tool orchestration	Workflow, coding, and QA task completion	Shows coordinated system behavior; does not measure validity of a fixed receiving agent after a controlled partner change.
Solver–verifier systems Cobbe et al. (2021); Lightman et al. (2024)	Solver and process or outcome verifier	Candidate scores, process labels, and verification feedback	Mathematical problem solving and verifier quality	Establishes verifier utility, not post-intervention validity under changed upstream outputs.
Learned judges: Prometheus Kim et al. (2024)	Generator and rubric-based evaluator	Rubrics and fine-grained evaluation scores	Evaluation capability and agreement with judgments	Studies judge quality on evaluation tasks, not calibration stability under a producer intervention.
Co-training: MAPoRL, WaltzRL, CoMAS Park et al. (2025); Zhang et al. (2026b); Xue et al. (2026)	Collaborating policies and conversation or feedback agents	Shared trajectories, joint updates, and interaction rewards	Collaboration, safety, and post-training performance	Studies co-adaptation; does not isolate an upstream-only change with a frozen downstream receiving agent.
Multi-agent RL and stability: simulator diversity and Dr. MAS Yu et al. (2026b); Feng et al. (2026a)	Agent policies and simulator or training controller	Shared rewards, simulator interaction, and policy normalization	Multi-agent policy stability and task performance	Does not test task-validity loss at a fixed receiving agent boundary under partner replacement.
Episodic and reflective memory: Generative Agents, Reflexion Park et al. (2023); Shinn et al. (2023)	Agent, reflection process, and memory store	Retrieved experiences, summaries, and verbal reflections	Long-horizon simulation and agent task improvement	Memory gains do not by themselves establish partner-specific recovery under matched drift.
Confidence-weighted reflection memory: Remember and Reweight Jin et al. (2026)	Debate agents and experience memory	Retrieved experience with confidence-based weighting	Multi-agent debate performance	Evaluates debate assistance, not a frozen receiving agent receiving changed-partner messages.
Memory evaluation and co-optimization: MemoryAgentBench, CoMAM Hu et al. (2026b); Mao et al. (2026)	Memory agent, extractor, profile builder, retriever, and receiving agent	Incremental retrieval and jointly optimized memory components	Memory competencies and memory-pipeline performance	Does not compare static and refreshed memory on identical changed-partner trajectories.
Communication and composition robustness Cheng et al. (2026); Hossain et al. (2026)	Sender, receiver, and safety evaluator	Relayed latent states and cross-agent composition	Causal communication and compositional safety	Focuses on channel causality or safety, not task validity of a fixed receiving agent after partner change.
Heterogeneous collaboration: RACER Lou et al. (2026)	Heterogeneous agents with routed roles	Cross-agent collaboration over knowledge-graph tasks	Explainable reasoning and collaboration quality	Cross-model collaboration is not a controlled replacement intervention on a frozen receiver.
In-context and cognitive interference Qian et al. (2026); Kliegl & Bäuml (2021)	One model processes conflicting or interfering information	Conflicting context and cognitive interference mechanisms	Selective in-context forgetting and cognitive evidence	The intervention is within a model or cognitive task, not a sender-receiver interface change.
Multi-agent drift and constraint preservation Xu et al. (2026b); Li et al. (2026); Rath (2026)	Handoffs, delegated agents, and long interaction trajectories	Authorization and safety state across multi-agent traces	Authorization preservation, constraint drift, and behavioral stability	Important drift neighbors, but outcomes differ from task-grounded validity of a frozen receiving agent.
Pipeline evaluation and baseline comparison: Two Calls Beat Five Agents Prajapati & Mohite (2026)	Multi-agent pipeline and single-model/self-refinement baselines	Matched task prompts and alternative inference pipelines	System-level task performance and efficiency	Compares systems as wholes; does not localize upstream changes at a validated receiving agent boundary.