SAME MESSAGE, SAME CREDIT? DIAGNOSING CREDIT STABILITY UNDER REROUTING IN MULTI-AGENT LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

In multi-agent question answering, determining whether a message's contribution remains valid after communication routes change is essential for reliable system redesign, because outdated scores can misidentify the evidence exchanges on which answers depend. Existing collaboration and attribution methods do not by themselves quantify the error of reusing contribution scores across communication graphs. Alternative routes can make an indispensable message redundant without changing its content. Structural summaries can also merge cases with different effects, preventing predictors from distinguishing their contributions. We propose a three-module diagnostic framework: paired measurement compares retained and disrupted messages on matched tasks, and cross-graph stability computes the optimal fixed credit and its exact worst-case error over evaluated graphs. Exposure sufficiency converts equal-summary effect differences into empirical worst-case error bounds and identifies the responsible states. Measured controlled credits give optimal fixed-score error 0.5 in all 12 settings, factorial equal-summary bounds are 0.4–0.75, and 160 questions from the 2WikiMultiHopQA multi-hop question answering dataset yield a statistically uncertain 0.0372 graph difference in Token-Overlap F-Measure (F1) (95% interval $[-0.0024, 0.0785]$). These diagnostics distinguish measuring a message's contribution from establishing its reuse, helping developers assess existing scores when communication patterns change.

1 INTRODUCTION

A multi-agent system uses Large Language Models (LLMs) to divide a task among agents that exchange evidence before a final agent responds. Answering a question may require facts found by different agents, so a missing message can prevent a correct answer. Changing these routes can reduce communication or provide alternative evidence, but may invalidate earlier assessments of which messages matter. Our task is to determine whether the same message's measured contribution to final-answer quality can be reused after communication routes change, and quantify the error when it cannot.

Existing research organizes agent roles, selects communication routes, and attributes outcomes to decisions or channels (Wu et al., 2024; Hong et al., 2024; Zhang et al., 2025b; Chen et al., 2026; Li et al., 2026). Optimizing communication or attributing an outcome does not by itself establish that a measured message credit remains accurate after the graph changes. We measure a message's *causal credit* through matched executions with it retained and deleted or replaced by conflicting content. Credit is intervention minus retained-message score; negative credit means disruption harms answer quality.

The difficulty begins with alternative routes. A message carrying the only copy of a required fact is indispensable, but a bypass can make it redundant. This *cross-graph credit instability* means one fixed score must misrepresent its effect in at least one graph. We seek the smallest worst-case error any reused score can achieve.

A coarse structural summary does not necessarily resolve this difficulty. A summary such as the number of disrupted downstream paths can merge different routes or missing-message combinations. When their effects differ, this *exposure collision* forces a summary-based predictor to treat unequal

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

Figure 1 follows the same evidence through four executions: adding a route changes whether deleting the message prevents a correct answer.

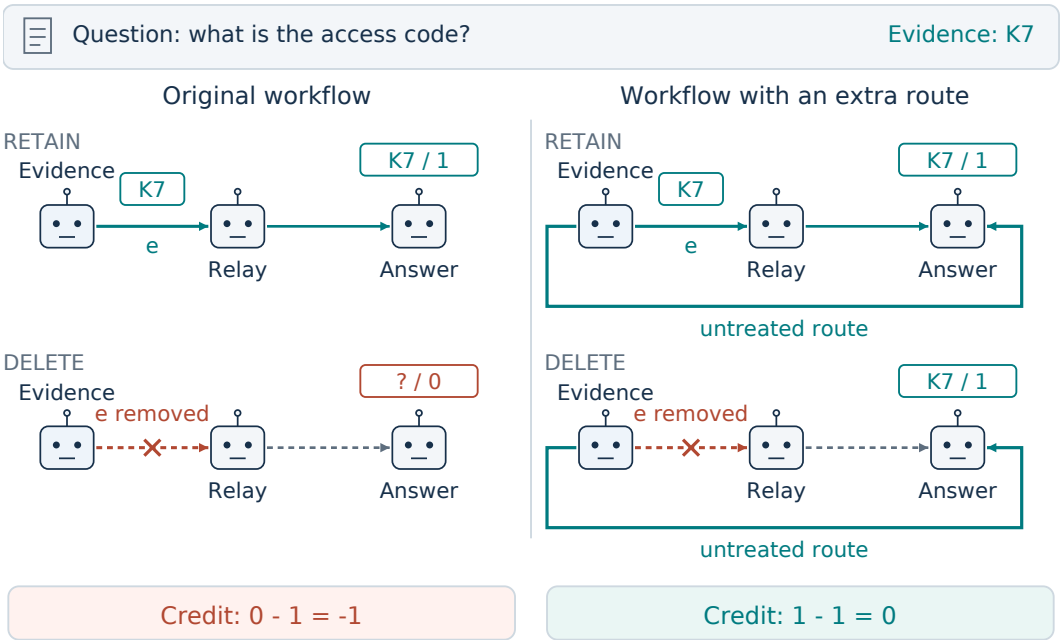

Figure 1: A worked illustration of the reuse decision. Output bubbles show answer / correctness; "?" denotes missing evidence. Only the focal message e is deleted, and the additional route remains intact. The scores illustrate the task, rather than an observed experimental trace.

contributions alike. Here, exposure means a pre-outcome summary of the graph and surrounding interventions; collisions bound the prediction error that remains when using it.

After measuring paired effects, our stability and sufficiency modules quantify variation across graphs and within summaries, respectively. They return error bounds and responsible cases on evaluated conditions; prediction on unobserved graphs requires additional evidence. Our contributions are:

- We propose a cross-graph stability diagnostic computing the optimal fixed credit and its exact worst-case approximation error over a finite set of evaluated graphs.
- We design an exposure sufficiency diagnostic that bounds empirical worst-case error from equal-summary effect differences and identifies states the summary cannot distinguish.
- We design a paired evaluation protocol combining controlled graph changes, factorial interventions, and natural question answering to test both diagnostics and reconstruction of held-out graph means.

2 Related Work

2.1 Multi-agent collaboration and communication design

Multi-agent LLM systems divide tasks among communicating agents and specify their roles, interaction procedures, and routes. Li et al. (2023) proposed Communicative Agents For "Mind" Exploration Of Large Language Model Society (CAMEL), which uses role-playing and inception prompts to guide collaborative task completion. Wu et al. (2024) developed the AutoGen conversation framework, which combines language models, tools, and human input through programmable conversations. Hong et al. (2024) proposed the MetaGPT meta-programming framework, which encodes standard operating procedures into prompts for specialized agents to produce and verify intermediate work. Zhang et al. (2025b) proposed G-Designer, which uses a variational graph autoencoder with agent and task nodes to generate task-specific communication graphs. Liu et al. (2024)

proposed the Dynamic LLM-Powered Agent Network (DyLAN), which selects agents through contribution estimates and adjusts their interactions across rounds. Zhang et al. (2025a) proposed AgentPrune, which learns communication masks and prunes redundant spatial and temporal connections. Wang et al. (2025) proposed AgentDropout, which optimizes round-specific adjacency matrices to remove redundant agents and exchanges. These methods configure participation and information flow through role assignment, graph generation, or removal of agents and links.

2.2 Credit assignment and failure attribution

Credit assignment and failure attribution connect outcomes to individual actions, messages, or execution steps using histories and counterfactual comparisons. Foerster et al. (2018) proposed Counterfactual Multi-Agent (COMA) Policy Gradients, whose centralized critic marginalizes one agent's action while holding the others fixed to compute an agent-specific advantage. Chen et al. (2026) proposed Contextual Counterfactual Credit Assignment (C3), which restores decision histories, samples alternative actions under a frozen policy, and uses a leave-one-out baseline for per-decision advantages. Li et al. (2026) proposed the E2-Explainer topology explanation framework, which masks channels and measures changes in outcomes and response stability to select explanatory subgraphs. Zhang et al. (2025c) introduced Who&When, a benchmark evaluating predictions of the responsible agent and failure step. Wang et al. (2026) proposed Causal Hierarchical Failure Attribution (CHIEF), which builds hierarchical causal graphs and combines oracle-guided backtracking with counterfactual screening. The resulting explanations concern action advantages, communication subgraphs, or failure locations within the evaluated executions.

2.3 Interference and effect transport

Causal inference under interference studies outcomes affected by other units' assignments; effect transport studies when causal effects can be reused across environments. Aronow & Samii (2017) developed a framework combining assignment designs, exposure mappings, and inverse-probability estimators. Ugander et al. (2013) proposed graph cluster randomization, which assigns treatments by cluster and computes network-exposure probabilities for weighted estimation. Sävje (2024) separated exposure-effect definitions from assumptions about complete interference, giving conditions for estimation with misspecified mappings. Leung (2022) allowed interference to decay with network distance and established consistency and asymptotic normality under approximate neighborhood interference. Pearl & Bareinboim (2014) developed a framework using selection diagrams and do-calculus to identify effects across populations. These studies specify the assignment process, information retained by exposure, and assumptions relating effects across states or populations.

3 Method

Our three-module framework first measures paired message effects, then diagnoses fixed-score reuse and the sufficiency of a structural summary. The paired-intervention measurement module runs matched tasks with the message retained and intervened upon, producing a table of effects. The cross-graph stability module averages these effects within each graph and compares the extreme credits, returning the best fixed score, its error, and the two responsible graphs. The exposure sufficiency module groups the same effects by a prespecified structural summary and compares effects within each group, returning error bounds and states that the summary cannot distinguish.

3.1 Task and message credit

We assess a focal message event e at scored node v across directed communication graphs G, holding the model, task distribution, prompting protocol, intervention, and scorer fixed. Let $Y_v(z_e, \mathbf{z}_{-e}; G)$ be the potential outcome: $z_e = 0$ retains the message, $z_e = 1$ deletes or replaces it, and $\mathbf{z}_{-e}$ specifies interventions on other messages. At state $s = (G, \mathbf{z}_{-e})$, credit is

$$C_{e,v}(G, \mathbf{z}_{-e}) = \mathbb{E}[Y_v(1, \mathbf{z}_{-e}; G) - Y_v(0, \mathbf{z}_{-e}; G)]\,. \tag{1}$$

The expectation covers task instances and generation randomness; negative credit means harm from intervention. Credit depends on both the intervention and the scored node: deleting a fact may make

Figure 2 shows the framework. Natural-task scores use Token-Overlap F-Measure (F1) and Exact Match (EM).

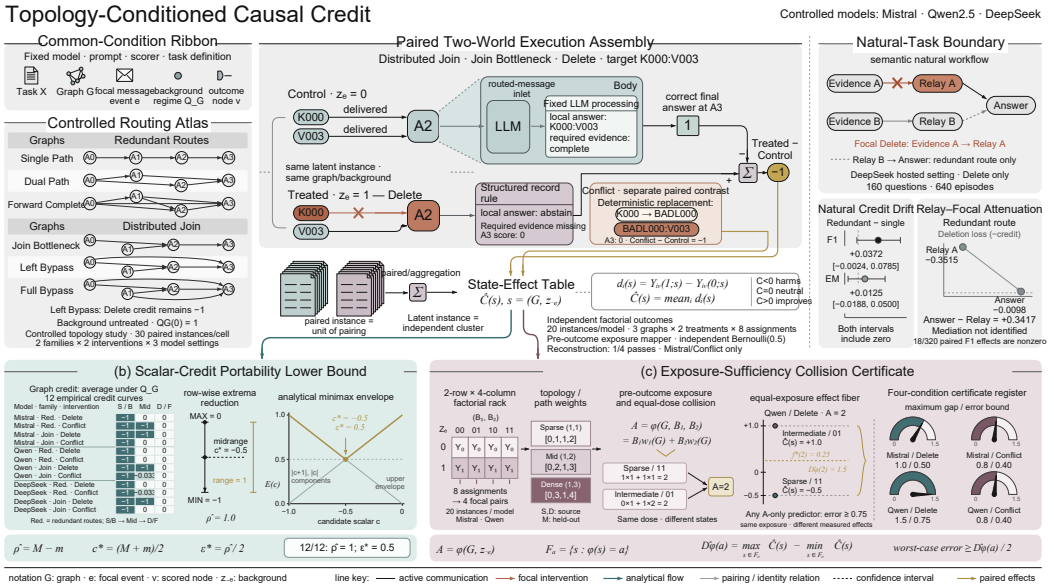

Figure 2: Complete diagnostic framework. Graph labels denote Sparse (S), Bottleneck (B), Intermediate (M/Mid), Dense (D), and Full Bypass (F); other labels denote Redundant Routes (Red.), Minimum (MIN), and Maximum (MAX). A0–A3 denote Agents 0–3. The relay panel plots deletion loss ($-C$), with point labels reporting credit (C). Exposure is specified before inspecting outcomes.

a relay abstain while a final answerer recovers it through a bypass. Removing a misleading message can instead improve the answer. For declared background weights Q_G, graph credit and its shift are

$$C^Q_{e,v}(G) = \mathbb{E}_{\mathbf{Z}_{-e}\sim Q_G}\left[C_{e,v}(G, \mathbf{Z}_{-e})\right], \tag{2}$$

$$\Gamma_e(G_A, G_B) = C^Q_{e,v}(G_B) - C^Q_{e,v}(G_A). \tag{3}$$

Thus the shift compares intervention effects, not unperturbed scores. The topology and natural question-answering studies set $Q_G(\mathbf{0}) = 1$; the factorial study averages four equiprobable background states under independent Bernoulli(0.5) assignments.

3.2 Paired-intervention measurement

This module isolates the focal-message effect through aligned tasks and surrounding conditions. For task i at state s, the paired difference and state-level estimate are

$$d_i(s) = Y_{iv}(1, \mathbf{z}_{-e}; G) - Y_{iv}(0, \mathbf{z}_{-e}; G), \qquad \widehat{C}(s) = \frac{1}{n_s}\sum_{i=1}^{n_s} d_i(s). \tag{4}$$

Both arms are executed for every positive-weight background state. For each graph and background assignment, pairing holds the task instance, scorer, focal event, and nonfocal interventions fixed; descendants execute using the evidence actually delivered in that arm. Delete removes the focal payload, whereas Conflict substitutes an incorrect value specified independently of the outcome. The released implementation preserves event identifiers and original payloads, records the observed payload separately, and checks a stored replacement hash before applying Conflict. This makes the edited event auditable without treating unrelated downstream responses as direct interventions.

The analysis code first averages repeated outcomes within each task–arm cell and then subtracts the two arm means. It emits an effect only when both arms exist; a missing arm must not be interpreted as a zero effect. Our reported paired studies contain complete pairs. For uncertainty, an entire task cluster is resampled with all its graph and assignment records intact, preserving the dependencies used in contrasts. The resulting effect table is the common input to both diagnostics; below, e, v are suppressed where unambiguous.

3.3 CROSS-GRAPH STABILITY DIAGNOSTIC

This module quantifies error from one score representing all evaluated graphs. For graph set $\mathcal{G}$, their range is

$$\widehat{\rho}_e = \max_{G\in\mathcal{G}} \widehat{C}^Q_{e,v}(G) - \min_{G\in\mathcal{G}} \widehat{C}^Q_{e,v}(G). \tag{5}$$

Proposition 1 (fixed-score approximation). On a finite, nonempty set of measured graph credits, the optimal fixed score and its error are

$$\min_{c\in\mathbb{R}} \max_{G\in\mathcal{G}} \left|\widehat{C}^Q_{e,v}(G) - c\right| = \frac{\widehat{\rho}_e}{2}, \qquad c^*_e = \frac{\max_G \widehat{C}^Q_{e,v}(G) + \min_G \widehat{C}^Q_{e,v}(G)}{2}. \tag{6}$$

To see why this is exact, write the smallest and largest supplied credits as m and M. For every c, the triangle inequality gives $M - m \le |M - c| + |c - m| \le 2\max_G |\widehat{C}^Q(G) - c|$. The midpoint $(m + M)/2$ is at distance at most $(M - m)/2$ from every credit, proving both the lower bound and attainability. Returning the two extreme graphs makes the error actionable: a zero range permits exact reuse on this set, whereas a positive range identifies two conditions that no fixed score can fit simultaneously.

3.4 EXPOSURE SUFFICIENCY DIAGNOSTIC

This module tests whether a structural summary merges states requiring different scores. The exposure, computed before observing outcomes, is

$$A = \phi(G, \mathbf{Z}_{-e}). \tag{7}$$

For the finite state set $\mathcal{S}$ and each observed summary value a, define

$$\mathcal{F}_a = \{s \in \mathcal{S} : \phi(s) = a\}, \qquad \widehat{D}_\phi(a) = \max_{s\in\mathcal{F}_a} \widehat{C}(s) - \min_{s\in\mathcal{F}_a} \widehat{C}(s). \tag{8}$$

States with equal summaries but different effects form an exposure collision. *Proposition 2 (summary-limited approximation).* For each nonempty group, every predictor using only A satisfies

$$\max_{s\in\mathcal{F}_a} \left|\widehat{C}(s) - f(a)\right| \ge \frac{\widehat{D}_\phi(a)}{2}. \tag{9}$$

The bound concerns estimated state-specific effects; averaging states may cancel their differences. Proposition 2 follows by applying the preceding argument inside each group $\mathcal{F}_a$. In fact, the bound is attainable on the supplied table: choosing the midpoint of the minimum and maximum effect separately for each observed a gives

$$\min_f \max_{s\in\mathcal{S}} |\widehat{C}(s) - f(\phi(s))| = \frac{1}{2} \max_a \widehat{D}_\phi(a). \tag{10}$$

Thus the diagnostic measures a limitation of the information given to a predictor, independently of its training algorithm. If a richer summary ψ retains ϕ, so that $\phi = h \circ \psi$, each ψ-group is contained in a ϕ-group and its diameter cannot increase. Refinement can therefore reduce this empirical error floor, although smaller groups may be harder to estimate and the inequality does not guarantee improved prediction on new tasks. The returned extreme states supply concrete distinctions to consider when refining the summary.

3.5 CONDITIONS FOR REUSE AND ESTIMATION UNCERTAINTY

This step separates exact diagnostics on measured effects from assumptions needed to predict another graph. Let $\delta^Q_G(a)$ average focal effects within exposure a under graph G and regime Q_G. With the task distribution fixed, if every target exposure occurs in the source and $\delta^Q_G(a) = \delta^Q_{G'}(a)$ for every target value, then

$$C^Q_{e,v}(G') = \sum_a \delta^Q_G(a) P^Q_{G'}(A = a). \tag{11}$$

Overlap makes the source terms available; conditional-effect equality justifies reuse. Reconstruction pools source effects within each summary value and weights them by the target assignment probabilities. An unseen target value requires an additional extrapolation rule. The identity concerns a graph average, so state-level errors can cancel despite exposure collisions. Section 4.3 tests whether conditioning improves this target prediction relative to an unconditioned source mean.

These approximation errors are distinct from estimation uncertainty. If every supplied state-effect estimate differs from its population effect by at most ϵ, the estimated half-diameter differs from its population counterpart by at most ϵ; the graph half-range obeys the same bound when its credits satisfy that uniform-error condition. This follows because a minimum or maximum changes by at most ϵ under uniform perturbation. Our task-cluster bootstrap intervals concern specified effects and contrasts, not a simultaneous uniform-error event; the largest-collision interval does not adjust for selecting an extreme. The guarantees therefore describe the supplied effect table; claims about population effects or unmeasured graphs require additional evidence.

4 Experiments

We test fixed-score reuse, path-summary sufficiency, and prediction across controlled and natural tasks. Controlled rerouting isolates alternative evidence paths, factorial assignments test information lost by a summary, and natural questions assess these diagnostics with imperfect answers. Bounds describe the supplied estimates; paired cluster-bootstrap Confidence Intervals (CIs) quantify sampling uncertainty separately.

4.1 Fixed scores across controlled communication graphs

We test whether rerouting makes a fixed message score inaccurate using two four-node tasks: opaque-code recovery (redundant routes) and complementary-atom combination (distributed join), each under three graphs. Delete removes the focal payload; Conflict replaces its value while retaining sender, recipient, fact identity, role, and template. Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Qwen2.5-7B-Instruct (Qwen Team, 2024), and a fixed DeepSeek hosted setting each use 30 matched latent instances per condition. Nodes abstain when required evidence is missing; correctness requires a valid answer matching private gold. The opaque values prevent task knowledge from supplying a missing answer: redundant routes carry the same code, whereas a join needs both a left and a right atom. Retained-message correctness is 1 in every topology cell. Local models use greedy decoding and the hosted setting uses temperature zero; prompts and instance seeds are aligned across arms and graphs.

Table 1: Full controlled credit curves, ordered as sparse/bottleneck $\rightarrow$ intermediate $\rightarrow$ dense/full bypass. Each entry uses 30 paired instances. All 12 curves have range 1 and fixed-score error floor 0.5.

Model	Task	Delete credit	Conflict credit
Mistral	Redundant routes	$(-1.000, 0.000, 0.000)$	$(-1.000, -1.000, 0.000)$
	Distributed join	$(-1.000, -1.000, 0.000)$	$(-1.000, 0.000, 0.000)$
Qwen	Redundant routes	$(-1.000, 0.000, 0.000)$	$(-1.000, 0.000, 0.000)$
	Distributed join	$(-1.000, -1.000, 0.000)$	$(-1.000, -0.033, 0.000)$
DeepSeek	Redundant routes	$(-1.000, 0.000, 0.000)$	$(-1.000, -0.033, 0.000)$
	Distributed join	$(-1.000, -1.000, 0.000)$	$(-1.000, 0.000, 0.000)$

The endpoint credit increases by 1 in all 12 comparisons, with planned-contrast intervals $[1, 1]$; the best fixed score is -0.5 and its worst-case error is 0.5 (Table 1). In the intermediate redundant graph, Qwen and DeepSeek lose 1 and 0.967 less correctness than Mistral under Conflict. A left-only bypass leaves join deletion harm unchanged.

Route availability therefore does not alone determine recovery; the saturated endpoints reflect the controlled construction. The intermediate graphs are more informative about processing behavior: the same redundancy that eliminates Delete loss can leave Conflict loss intact. In the join task, a left-only bypass supplies only one of the two required atoms, explaining why deletion can remain

harmful until both inputs reach the answerer. These comparisons distinguish the existence of a route from delivery of sufficient, usable evidence.

4.2 Different effects under the same path summary

To test whether equal path summaries hide unequal effects, we execute all eight assignments to one focal and two background messages for 20 latent instances per graph and intervention. The pre-outcome summary excludes the focal intervention:

$$A = \phi(G, B_1, B_2) = B_1 w_1(G) + B_2 w_2(G), \tag{12}$$

where B_j marks disruption of background message j and w_j counts its downstream paths. Source weights are $(1, 1)$ and $(1, 3)$; the held-out graph has $(1, 2)$. Mistral and Qwen each contribute 960 episodes and 5,120 valid node calls across Delete and Conflict. Bootstrap clusters are latent instances.

Table 2: Maximum collisions. States are graph/background assignment; graph labels are Sparse (S), Intermediate (M), and Dense (D). The supplied summary A is identical within each row.

Model	Edit	A	Min state / effect	Max state / effect	Gap	Bound
Mistral	Delete	2	S/11: -1.000	M/01: 0.000	1.000	0.500
	Conflict	1	S/01: -0.800	D/10: 0.000	0.800	0.400
Qwen	Delete	2	S/11: -0.500	M/01: 1.000	1.500	0.750
	Conflict	1	S/01: -0.800	D/10: 0.000	0.800	0.400

Equal-summary effects differ by 0.8–1.5, forcing error at least 0.4–0.75 (Table 2). For Qwen Delete at dose 2, credit is -0.5 for sparse/11 and $+1$ for intermediate/01. The selected gap has descriptive interval $[1.3, 1.7]$, without selection adjustment.

This summary loses route and assignment information, irrespective of predictor training; richer summaries remain untested. The Qwen Delete witness also reverses the sign of credit: the same summary represents a harmful deletion in one state and a helpful deletion in another. No predictor receiving only A can fit both measured state effects; fitting them separately requires an input that distinguishes the states. The finding concerns this scalar mapping, rather than every possible graph representation.

4.3 Reconstructing the average effect in a held-out graph

We test whether conditioning on path dose improves graph-mean prediction by pooling factorial source effects within doses and weighting them by the held-out graph's four equiprobable background states. The naive comparator averages the two source graph means. The prespecified criterion requires full dose support, at least 20% relative error reduction, and an improvement interval above zero. All combinations have full support; intervals use 4,000 latent-instance bootstrap resamples, retaining all assignments. Joint success requires both interventions to pass for both models.

Table 3: Absolute errors for the held-out graph mean. Improvement is naive minus reconstruction error; relative reduction is zero when the initial error is zero.

Model	Intervention	Naive error	Reconstruction error	Relative reduction	Improvement 95% CI
Mistral	Delete	0.0000	0.0000	0.000	$[0.0000, 0.0000]$
Mistral	Conflict	0.1938	0.1292	0.333	$[0.0542, 0.0750]$
Qwen	Delete	0.1313	0.0958	0.270	$[-0.0042, 0.0688]$
Qwen	Conflict	0.0875	0.1208	-0.381	$[-0.0396, -0.0271]$

Only Mistral Conflict passes: its error falls by approximately 0.0646 (33.3%). Qwen Delete improves by about 0.035 (27.0%), but its interval includes zero; Qwen Conflict worsens by 0.0333 (Table 3). Thus coverage permits computation without ensuring transfer. Zero Mistral Delete mean error can coexist with a nonzero collision bound because state-level errors cancel on averaging.

4.4 FINAL-ANSWER EFFECTS IN NATURAL QUESTION ANSWERING

We test whether routing changes final-answer message credit on 160 preselected 2WikiMultiHopQA validation questions balanced across four types (Ho et al., 2020). Evidence agents communicate through relays to an answerer. Delete removes the primary evidence message; an untreated auxiliary relay reaches the answerer only in the redundant graph. Two graphs and both arms produce 640 complete episodes in one fixed DeepSeek hosted setting, with no selective retries. F1, the harmonic mean of token precision and recall, is primary; EM tests scoring sensitivity (Rajpurkar et al., 2016); bootstrap clusters are questions.

Table 4: Natural question-answering credit at the final answerer. Drift is redundant minus single; intervals resample paired questions.

Metric	Single	Redundant	Drift [95% CI]
F1	-0.0470	-0.0098	$0.0372\ [-0.0024, 0.0785]$
Exact match	-0.0188	-0.0063	$0.0125\ [-0.0188, 0.0500]$
Exploratory F1 strata (40 questions each)			
Bridge-comparison	-0.0250	0.0000	$0.0250\ [-0.0500, 0.1000]$
Comparison	0.0500	0.0000	$-0.0500\ [-0.1250, 0.0000]$
Compositional	-0.1215	0.0000	$0.1215\ [0.0382, 0.2211]$
Inference	-0.0917	-0.0393	$0.0524\ [-0.0155, 0.1292]$

Redundancy reduces estimated deletion loss by 0.0372 F1, while the exact-match difference is 0.0247 smaller; both intervals include zero (Table 4). Only 18 of 320 question–graph F1 effects are nonzero. Control F1 is already zero for 86.9% of single-route questions and 55.6% under redundancy, restricting observable harm.

The evidence is consistent with reduced message dependence but does not establish a nonzero aggregate interaction. Question-type differences help interpret this average: the compositional drift exceeds the comparison drift by 0.1715 F1, and only the compositional interval excludes zero. These exploratory strata do not establish a confirmed between-type interaction, but show why a graph average should not be read as a common effect for every question.

4.5 RELAY EFFECTS AND FINAL-ANSWER EFFECTS

We test whether relay credit can substitute for final-answer credit using both outputs from every existing episode. For each scored node, we estimate Delete-minus-Control F1 and bootstrap the final-minus-relay difference by question, retaining the paired outputs from both routing conditions. This reuses all 640 episodes without new treatments or selection.

Table 5: Node-specific F1 credit and its 95% question-cluster interval. Final-minus-relay credit measures attenuation of deletion harm.

Topology	Relay	Final answer	Final − relay
Single	-0.0814	-0.0470	0.0344
	$[-0.1223, -0.0448]$	$[-0.0893, -0.0060]$	$[0.0094, 0.0656]$
Redundant	-0.3515	-0.0098	0.3417
	$[-0.4263, -0.2775]$	$[-0.0259, 0.0000]$	$[0.2675, 0.4171]$

Final-answer loss is 0.0344 F1 smaller than relay loss under single routing and 0.3417 smaller under redundancy; both difference intervals exclude zero (Table 5). Redundancy increases this attenuation by 0.3073 F1 in the point estimates. Relay credit therefore cannot substitute for final-answer credit. An untreated route is consistent with the larger attenuation, but this contrast does not distinguish aggregation, selective evidence use, and repair.

4.6 PREDICTING CREDIT FROM STRUCTURAL INFORMATION

We test whether structural context improves prediction of 320 question–graph F1 effects, excluding every effect of the held-out question from fitting in leave-one-question-out evaluation. Holding out both graph records prevents the same question's other measured effect from entering the training means. Comparators use pooled or topology-specific means, outcome-free connectivity, and question type. The six descriptors are reachability, surviving path count, edge-disjoint paths, minimum cut, shortest-path disconnection, and untreated bypass count; each coincides with topology here. Reachability records whether an untreated route survives; surviving simple-path count, edge-disjoint path count, and minimum cut describe surviving source–sink connectivity. Shortest-path disconnection records loss of the available path, and untreated bypass count records routes avoiding the treated edge. Degree and density have no additional variation and supply no separate predictor comparison. The deterministic necessity oracle predicts binary harm without fitting. We compare Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

Table 6: Held-out prediction errors. All six connectivity descriptors give the topology row's errors. The oracle is a binary comparator, not a fitted magnitude model.

Predictor	MAE	RMSE
Pooled scalar	0.07321	0.20282
Topology / each connectivity descriptor	0.07212	0.20251
Structure + question type	0.08440	0.20116
Structure $\times$ question type	0.08130	0.20039
Deterministic necessity oracle	0.48139	0.70328

Topology reduces MAE by 0.00109 (1.5%) and RMSE by 0.00031 relative to pooling. Adding type increases MAE by 0.01228, or 0.00918 with interactions, despite small RMSE reductions (Table 6). Coincident descriptors cannot isolate structural mechanisms, and these small gains do not establish predictive value on new graphs.

The controlled tasks deliberately saturate correctness. Correctness requires schema validity, delivery validity, and exact equality to the scorer-private gold; missing required evidence forces abstention. The gold answer is not separately revealed in the prompt. Delete removes the full payload without padding or length matching, whereas Conflict substitutes its content. Control and Conflict are not exact-token matched, leaving lexical and tokenization differences as possible confounds.

The natural study covers one dataset, one hosted setting, two small graphs, and deletion. The hosted DeepSeek label has no public checkpoint mapping. Our diagnostics assess information loss rather than a pruning policy. The work uses benchmark traces without human subjects; protocols and independent bootstrap units are specified, and all 160 selected questions are retained.

5 CONCLUSION

For multi-agent LLMs exchanging evidence through changing routes, we study message-score reuse and provide diagnostics that quantify its error. Existing collaboration and attribution methods do not by themselves quantify the error of transferring a measured effect to another configuration. An alternative route can change the same message's effect, making one fixed score inaccurate across graphs. Even with structural context, a summary can merge states with different effects and hide the distinction a predictor needs. We therefore propose a cross-graph stability diagnostic that compares graph-credit extremes to compute the best fixed score and its exact worst-case error over the evaluated graphs. Our exposure sufficiency diagnostic converts equal-summary gaps into empirical worst-case error bounds; paired measurement supplies both modules, which return bounds and responsible graph or state pairs. All 12 measured controlled credit surfaces have optimal fixed-score error 0.5, factorial equal-summary bounds are 0.4–0.75, and 160 2WikiMultiHopQA questions yield a statistically uncertain graph difference of 0.0372 F1 (95% interval $[-0.0024, 0.0785]$). A remaining limitation is testing only one scalar path-count summary, leaving richer representations unexamined. Guided by the $d/2$ error bound for merging effects separated by d, we plan to add pre-outcome route identity and surviving-route information to reduce collisions and test held-out prediction.

REFERENCES

Peter M. Aronow and Cyrus Samii. Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics*, 11(4):1912–1947, 2017. doi: 10.1214/16-AOAS1005.

Yanjun Chen, Yirong Sun, Hanlin Wang, Jinghan Wang, Xinming Zhang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Exact is easier: Credit assignment for cooperative LLM agents. arXiv preprint arXiv:2603.06859, 2026.

DeepSeek-AI. DeepSeek-V3 technical report. arXiv preprint arXiv:2412.19437, 2024.

Falcon Team. Welcome to the Falcon 3 family of open models! Technology Innovation Institute, 2024. Model-family release, December 17, 2024.

Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 2974–2982, 2018. doi: 10.1609/aaai.v32i1.11794.

Jie He, Nan Hu, Wanqiu Long, Jiaoyan Chen, and Jeff Z. Pan. MINTQA: A multi-hop question answering benchmark for evaluating LLMs on new and long-tail knowledge. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 445–479. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.18.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.580.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*, volume 2024, pp. 23247–23275, 2024.

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. arXiv preprint arXiv:2310.06825, 2023.

Timothy Kassis, Vinayak Agarwal, Yuhuan He, Darshil Patel, and Aubrey M. Brueckner. Scientific agent skills: A library of procedural knowledge for research agents. arXiv preprint arXiv:2609.00065, 2026.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling*, 2025.

Michael P. Leung. Causal inference under approximate neighborhood interference. *Econometrica*, 90(1):267–293, 2022. doi: 10.3982/ECTA17841.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008, 2023. doi: 10.52202/075280-2264.

Junzhi Li, Peng He, Qirui Ji, Wei Wang, Lixiang Liu, and Chuxiong Sun. Discovering efficient and explainable communication topologies for LLM-based multi-agent systems via causal inference. arXiv preprint arXiv:2608.12921, 2026.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. A dynamic LLM-powered agent network for task-oriented agent collaboration. In *First Conference on Language Modeling*, 2024.

Judea Pearl and Elias Bareinboim. External validity: From do-calculus to transportability across populations. *Statistical Science*, 29(4):579–595, 2014. doi: 10.1214/14-STS486.

Qwen Team. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115, 2024.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1264.

Fredrik Sävje. Causal inference with misspecified exposure mappings: separating definitions and assumptions. *Biometrika*, 111(1):1–15, 2024. doi: 10.1093/biomet/asad019.

Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. Graph cluster randomization: Network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 329–337. Association for Computing Machinery, 2013. doi: 10.1145/2487575.2487695.

Yawen Wang, Wenjie Wu, Junjie Wang, and Qing Wang. From flat logs to causal graphs: Hierarchical failure attribution for LLM-based multi-agent systems. arXiv preprint arXiv:2602.23701, 2026.

Zhexuan Wang, Yutong Wang, Xuebo Liu, Liang Ding, Miao Zhang, Jie Liu, and Min Zhang. AgentDropout: Dynamic agent elimination for token-efficient and high-performance LLM-based multi-agent collaboration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24013–24035, Vienna, Austria, jul 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1170.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *International Conference on Learning Representations*, volume 2025, pp. 75389–75428, 2025a.

Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, Kun Wang, Tianlong Chen, and Dawei Cheng. G-Designer: Architecting multi-agent communication topologies via graph neural networks. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76678–76692. PMLR, 2025b.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, and Qingyun Wu. Which agent causes task failures and when? On automated failure attribution of LLM multi-agent systems. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76583–76599. PMLR, 13–19 Jul 2025c.

ARTIFICIAL INTELLIGENCE (AI) USE STATEMENT

Generative AI assisted writing, clarification of definitions and assumptions, diagnostic formulation from existing results, mathematical checks, citation verification, and LaTeX preparation. Scientific Agent Skills' citation-management workflow (Kassis et al., 2026) assisted bibliographic validation. No additional experimental observations were generated; experiment descriptions identify the evaluated LLM agents.

A SUPPLEMENTARY EXPERIMENTAL MATERIAL

This appendix records the controlled-task protocols, complete diagnostic results, and supporting natural-question analyses. The final three studies retain their original intervention units and outcomes and are reported separately from the matched graph comparisons.

A.1 CONTROLLED TASKS AND MESSAGE INTERVENTIONS

This protocol specifies how routing can change access to required evidence while the task and focal message are held fixed, making the fixed-score reuse test interpretable. Both task families use four nodes, A0–A3, and score A3's final answer. An edge ij denotes Ai→Aj.

Table A1: Controlled task graphs and focal message events.

Task family	Graph and edge set	Focal event
Redundant routes	Chain: $\{01, 12, 23\}$	01
	Dual path: $\{01, 02, 13, 23\}$	01
	Forward-complete: $\{01, 02, 03, 12, 13, 23\}$	01
Distributed join	Bottleneck: $\{02, 12, 23\}$	02
	Left-bypass: $\{02, 12, 23, 03\}$	02
	Full-bypass: $\{02, 12, 23, 03, 13\}$	02

Table A1 records the focal event and communication graphs for each family. In redundant routes, A0 sends an opaque code-valued fact through the focal event A0→A1, A1/A2 process the delivered evidence, and A3 must reproduce the exact code. In distributed join, A0 supplies a left-hand atom and A1 a complementary right-hand atom; A2 combines them into an ordered pair. The focal event is A0→A2, and the bypasses determine whether its left atom can still reach A3 after intervention.

Delete removes the full focal payload at the routing layer, without padding or length matching, so no focal atom is delivered through that edge. Conflict deterministically substitutes an incorrect opaque value while preserving sender, recipient, edge, message role, surface template, schema, and fact-identifier role. The replacement cannot equal the gold value and is fixed independently of outcomes; its meaning is unambiguous within these tasks. Control and Conflict are not exact-token matched, leaving lexical and tokenization differences as possible confounds.

The independent unit is the latent opaque-value instance. An index i deterministically specifies distinct gold and conflict values, and treatment assignment and replacement are fixed from the instance seed before execution. Within each task family, the question, fact-identifier vocabulary, focal event identifier, processing order, scorer, and decoding rule are frozen across arms and topologies. Delete changes both observed text and the observed atom list; Conflict changes observed text and status while retaining the focal atom identifier. These controls make the intervention and routing contrast reproducible while retaining the stated prompt-length limitation.

A.2 FULL CONTROLLED CREDIT SURFACES

We examine whether fixed-score reuse fails consistently across the evaluated models and task designs, using three model settings and the redundant-route and distributed-join tasks. Each of 30 matched latent instances per graph and condition is executed with the focal message retained, deleted, and replaced by a conflicting value across three communication graphs. Credit is intervention correctness minus retained-message correctness at the final node.

Full controlled credit matrix

30 paired instances per cell; credit = intervene - retain.

Model	Task	Intervention	Sparse / bottleneck	Intermediate	Dense / full bypass
Mistral	Redundant	Delete	-1.000	0.000	0.000
		Conflict	-1.000	-1.000	0.000
	Join	Delete	-1.000	-1.000	0.000
		Conflict	-1.000	0.000	0.000
Qwen	Redundant	Delete	-1.000	0.000	0.000
		Conflict	-1.000	0.000	0.000
	Join	Delete	-1.000	-1.000	0.000
		Conflict	-1.000	-0.033	0.000
DeepSeek	Redundant	Delete	-1.000	0.000	0.000
		Conflict	-1.000	-0.033	0.000
	Join	Delete	-1.000	-1.000	0.000
		Conflict	-1.000	0.000	0.000

Figure A1: Full controlled credit matrix. All $3 \times 2 \times 2 = 12$ model–family–treatment surfaces are shown over sparse/bottleneck, intermediate, and dense/full-bypass states. Every row has $\rho_e = 1$ and half-range lower bound 0.5. Structural state and topology identity coincide in this controlled design.

Figure A1 shows that moving from the bottleneck or sparse endpoint to the full-bypass or dense endpoint increases credit by 1.0 in all 12 combinations, with endpoint-contrast intervals $[1, 1]$. The nondegenerate intermediate estimates are -0.033 for Qwen distributed-join Conflict and DeepSeek redundant-route Conflict, both with 95% intervals $[-0.10, 0]$; all other intermediate estimates are exactly -1 or 0. Thus the best fixed score still incurs a worst-case error of 0.5 on each evaluated graph set, while the intermediate responses can differ across models. This separates a shared limitation of fixed scores from model-specific responses to added routes.

A.3 Controlled Execution and Worked Traces

We inspect paired executions to show how the specified message edits reach the scored answer through local evidence processing. Both focal arms use the same instance, model, graph, prompt, scorer, and processing order. The direct recipient is A1 in redundant routes and A2 in distributed join; A3 is the downstream scored node. Each processing node passes only its local answer, provenance over delivered facts, and confidence. Missing required evidence forces abstention, and correctness requires schema validity, delivery validity, and exact equality to the scorer-private gold. The prompt does not separately reveal the scorer-private gold answer. A deterministic parser may remove exactly one complete outer JavaScript Object Notation (JSON) Markdown fence labelled `json`, but cannot repair content.

The local Mistral and Qwen models use greedy decoding with at most 64 new tokens. Each contributes 720 complete traces from 1,800 valid node generations, and the hosted DeepSeek topology replication also contains 1,800 completed generations. The DeepSeek-V3 report (DeepSeek-AI, 2024) supplies model-family background without identifying the hosted checkpoint. Control correctness is 1.0 in every controlled topology cell.

Table A2: Worked controlled traces for instance $i = 0$. The gold values are `ZX-0000-011` and `K000:V003`, respectively.

Task	Arm	Local processing result	A3 outcome
Chain	Control	A1/A2 pass `ZX-0000-011`	correct
Chain	Delete	A1/A2 abstain: insufficient evidence	incorrect
Chain	Conflict	A1/A2 pass `QY-0000-019`	incorrect
Join	Control	A2 forms `K000:V003`	correct
Join	Delete	A2 abstains: left atom absent	incorrect
Join	Conflict	A2 forms `BADL000:V003`	incorrect

Table A2 gives the index-0 traces for the two bottleneck cases. For the redundant-route chain, A1 and A2 reproduce the correct code in Control, abstain after Delete, and propagate the substituted code after Conflict. For the distributed join, A2 combines the correct left atom with A1's right atom in Control, abstains without the left atom after Delete, and combines the incorrect substituted atom after Conflict. Both interventions therefore change each illustrated sink from correct to incorrect, producing a paired correctness effect of -1. The examples show how route-dependent evidence availability enters the scored outcome through the specified local processing rules.

A.4 Additional Fixed-Graph Propagation Checks

We check whether a local message effect persists one hop downstream by comparing direct-recipient and downstream correctness in fixed redundant-consensus and distributed-join workflows. Mistral and Qwen use 60 paired instances per family–intervention condition, while DeepSeek uses 30. The graph remains fixed within each check, and the effect at each readout is the intervention score minus the retained-message score.

Table A3: Controlled direct / one-hop effects. Local models use 60 paired instances per family–treatment condition; DeepSeek uses 30.

Family	Treatment	Mistral	Qwen2.5	DeepSeek
Redundant consensus	Delete	$0.000/0.000$	$0.000/0.000$	$0.000/0.000$
Redundant consensus	Conflict	$-1.000/-1.000$	$-0.783/-0.783$	$-0.900/-0.900$
Distributed join	Delete	$-1.000/-1.000$	$-1.000/-1.000$	$-1.000/-1.000$
Distributed join	Conflict	$-1.000/-1.000$	$-1.000/-1.000$	$-1.000/-1.000$

Table A3 shows that the two readouts have identical mean effects in every reported cell: the measured change from the direct recipient to one hop downstream is zero. For redundant consensus, Conflict reduces correctness by 0.783–1.0 more than Delete; the two interventions have the same mean effect for distributed join. Retained alternative evidence makes deletion harmless in the redundant-consensus workflow, whereas conflicting content remains disruptive. These results verify that the specified downstream readout registers the controlled intervention effect, but a fixed graph does not provide a test of portability across topologies.

A.5 Hosted DeepSeek Configuration and Reproducibility

This configuration record makes the hosted replication auditable by documenting the model identity and settings held fixed across controlled graphs. Calls use the provider's Application Programming Interface (API); timestamps use Coordinated Universal Time (UTC).

Table A4: Configuration used for the hosted DeepSeek controlled-topology replication.

Field	Recorded value
Provider	DeepSeek official API
Hosted model label	DeepSeek-Flash
Runtime identity	one stable provider-reported fingerprint
Decoding	reasoning disabled; temperature 0
Response mode	structured response
Output ceiling / timeout	1,024 tokens / 90 seconds
Execution window (UTC)	2026-09-15 16:35:31–16:54:22
Retry policy	no retry after a completed response
Identity consistency	1,800/1,800 generations share one model identity
Wrapper normalization	none required in this replication
Completion	1,800/1,800 successful generations; 0 provider failures
Execution record	configuration and outputs archived for reproducibility

Table A4 records the hosted configuration for the controlled-topology replication. All 1,800 generations completed successfully under one stable provider-reported model identity, with no provider failures or wrapper normalization. The provider exposes a hosted label rather than a public checkpoint identifier; the archived execution record preserves the configuration and outputs, and credentials are excluded from released artifacts.

A.6 Equal-Summary States and Error Bounds

To identify information lost by the path summary, we group factorial graph/background states by their scalar path dose, fixed before target evaluation, and compare focal treated-minus-control effects within each group. These are post-evaluation diagnostics, not representation-fitting data.

Table A5: Selected collisions: equal scalar path dose, different assignments/topologies, and different focal treated-minus-control effects. Each state is written as graph/background assignment; Intermediate (mid) identifies the held-out graph.

Model / treatment	Dose	State A	Effect A	State B	Effect B
Qwen / Delete	2	mid/01	1.000	sparse/11	-0.500
Mistral / Delete	2	mid/01	0.000	sparse/11	-1.000
Mistral / Delete	3	mid/11	-1.000	dense/01	0.000
Qwen / Delete	3	mid/11	-0.400	dense/01	0.450
Mistral / Conflict	1	mid/10	0.000	sparse/01	-0.800

Table A5 gives concrete collisions. The largest Qwen Delete gap exceeds the largest Mistral Delete gap by 0.5, increasing the error lower bound by 0.25. The maximum gap/bound pairs are $1.0/0.5$ for Mistral Delete, $0.8/0.4$ for Mistral Conflict, $1.5/0.75$ for Qwen Delete, and $0.8/0.4$ for Qwen Conflict. Equal path dose therefore merges states needing different effect predictions in these measurements. These bounds concern the supplied summary and estimated state effects; each task family has only one focal edge, so the study cannot assess ranking stability among multiple comparable edges.

A.6.1 Complete equal-dose groups

We inspect every observed equal-dose group containing multiple states to determine where the summary loses effect information and where it introduces no measured approximation error. This extends the selected examples to the full set of 16 eligible model–treatment–dose groups. Within each group, we identify the minimum- and maximum-effect states and compute their gap $D_\phi(A)$ and half-gap bound $D_\phi(A)/2$.

Table A6: Observed equal-dose groups, their extreme states, and resulting error bounds. Intermediate (mid) identifies the held-out graph.

Model / treatment	A	Min state: effect	Max state: effect	Gap	Bound
Mistral / Delete	0	sparse/00: 0.000	dense/00: 0.000	0.000	0.000
	1	sparse/01: 0.000	dense/10: 0.000	0.000	0.000
	2	sparse/11: −1.000	mid/01: 0.000	1.000	0.500
	3	mid/11: −1.000	dense/01: 0.000	1.000	0.500
Mistral / Conflict	0	sparse/00: 0.000	dense/00: 0.000	0.000	0.000
	1	sparse/01: −0.800	dense/10: 0.000	0.800	0.400
	2	sparse/11: 0.000	mid/01: 0.000	0.000	0.000
	3	mid/11: 0.000	dense/01: 0.000	0.000	0.000
Qwen / Delete	0	sparse/00: 0.000	dense/00: 0.000	0.000	0.000
	1	sparse/10: 0.000	sparse/01: 0.800	0.800	0.400
	2	sparse/11: −0.500	mid/01: 1.000	1.500	0.750
	3	mid/11: −0.400	dense/01: 0.450	0.850	0.425
Qwen / Conflict	0	sparse/00: 0.000	dense/00: 0.000	0.000	0.000
	1	sparse/01: −0.800	dense/10: 0.000	0.800	0.400
	2	mid/01: −0.700	sparse/11: 0.000	0.700	0.350
	3	mid/11: 0.000	dense/01: 0.000	0.000	0.000

Table A6 retains all 16 model–treatment–dose groups containing at least two graph/assignment compositions. Eight groups have positive measured gaps and eight have zero gaps, so the summary loses effect information in only part of the evaluated state partition. A zero gap permits exact fitting of the supplied effects within that group; it does not establish equality of population effects or performance on unmeasured states.

A.7 Held-Out Target-Mean Reconstruction

We test whether source effects conditioned on exposure improve prediction of the mean focal effect in an intermediate held-out topology, using exhaustive paired execution as the target reference. The naive prediction equally averages the marginal focal effects from sparse and dense training topologies. Exposure reconstruction instead averages the training conditional effects $\widehat{\delta}(a)$ over the target's exact assignment distribution. Error is absolute prediction error on this single target topology, and improvement is naive error minus reconstruction error. We use the latent instance as the bootstrap cluster, retain all eight assignment configurations inside each cluster, and run 4,000 percentile-bootstrap repetitions. The criterion fixed before target evaluation requires complete target exposure support, relative error reduction of at least 0.20, and an improvement interval strictly above zero.

Table A7: Exposure-reconstruction metrics under the criterion fixed before target evaluation.

Model / treatment	Naive error	Reconstruction error	Relative reduction	Improvement 95% CI	Criterion
Mistral / Delete	0.0000	0.0000	0.000	$[0.0000, 0.0000]$	no error to reduce
Mistral / Conflict	0.1938	0.1292	0.333	$[0.0542, 0.0750]$	pass
Qwen2.5 / Delete	0.1313	0.0958	0.270	$[-0.0042, 0.0688]$	CI crosses zero
Qwen2.5 / Conflict	0.0875	0.1208	−0.381	$[-0.0396, -0.0271]$	worsens

Table A7 shows that reconstruction reduces Mistral Conflict error by approximately 0.0646, or 33.3%, relative to the source-average prediction. Qwen Delete improves by about 0.035 (27.0%), but its improvement interval includes zero; Qwen Conflict error increases by approximately 0.0333. Only one of four model–treatment combinations succeeds, so the criterion requiring both treatments

to succeed for both models is unmet. Complete exposure support makes the calculation possible without guaranteeing that conditional effects transfer. This test concerns the marginal target effect, where state-specific errors can cancel; an equal-summary collision therefore need not produce nonzero marginal error.

A.8 2WikiMultiHopQA Scoring Sensitivity and Floor Effects

We check whether the natural-task graph contrast depends on the scoring metric and whether already-incorrect answers limit the deletion harm the scores can register. These checks delimit the interpretation of the full-sample credit difference. We use the same 160 questions to compare F1 with exact match, tabulate Control/Delete score distributions, and count harmful, unchanged, and helpful paired effects within each topology. Confirmatory inference retains all questions.

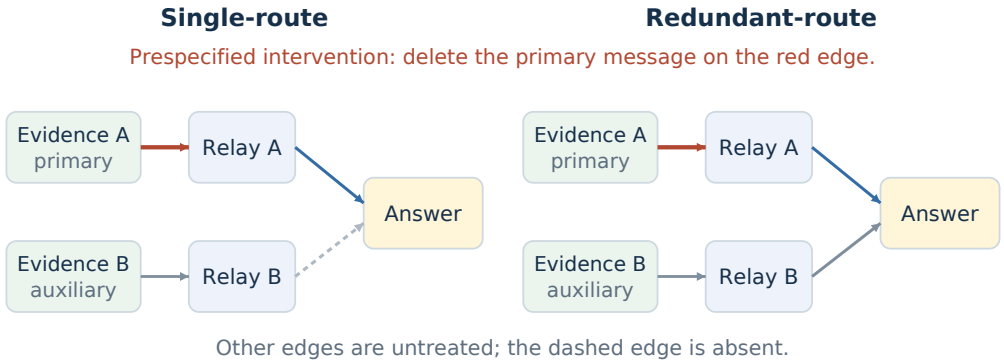

Figure A2: Natural-message routing design. The intervention removes the primary evidence message on the focal edge; the auxiliary route remains untreated.

Figure A2 specifies the routes: Delete removes the complete primary evidence message, while the auxiliary evidence–relay route remains untreated and reaches the final answerer only in the redundant graph.

Table A8: Primary F1 and exact-match sensitivity on the common credit scale; drift is redundant minus single.

Metric	Single credit	Redundant credit	Drift [95% CI]
F1	-0.0470	-0.0098	0.0372 $[-0.0024, 0.0785]$
Exact match	-0.0188	-0.0063	0.0125 $[-0.0188, 0.0500]$

Table A8 shows that the graph difference is 0.0247 smaller under exact match than F1; both intervals include zero. For F1, deletion increases the zero-score share by 6.25 percentage points under single routing (0.8688 to 0.9313), versus 1.25 points under redundancy (0.5563 to 0.5688). Median paired effects are zero for both metrics and topologies. Delete harms/leaves unchanged/helps 13/144/3 single-route and 2/158/0 redundant-route questions by F1; the exact-match counts are 6/151/3 and 1/159/0. Control F1 means are 0.1064 and 0.4170, medians are zero, and exact-one shares are 0.0625 and 0.3750. Many Control answers already receive the minimum score, limiting measured harm. Retaining these cases preserves the full-sample target; normalized harm is omitted because zero-Control cases would need an arbitrary stabilizer.

A.9 2WikiMultiHopQA Type-Stratified Results

To assess whether the aggregate graph contrast conceals question-type variation, we estimate the redundant-minus-single F1 credit difference separately for four benchmark types. Each stratum contains 40 questions from the existing 160-question 2WikiMultiHopQA sample and reuses their

matched executions. Uncertainty intervals resample question clusters, and the type analysis is exploratory.

Table A9: Exploratory type strata on the common credit scale. Drift is redundant-route credit minus single-route credit.

Type	Single credit	Redundant credit	Drift [95% CI]
Bridge-comparison	-0.0250	0.0000	0.0250 $[-0.0500, 0.1000]$
Comparison	0.0500	0.0000	-0.0500 $[-0.1250, 0.0000]$
Compositional	-0.1215	0.0000	0.1215 $[0.0382, 0.2211]$
Inference	-0.0917	-0.0393	0.0524 $[-0.0155, 0.1292]$

Table A9 shows that the estimated graph difference is 0.1715 larger for compositional than comparison questions. Only the compositional stratum's displayed interval excludes zero. The differences indicate that aggregation can conceal variation among question types, but these exploratory strata do not establish a confirmed between-type interaction or replace the full-sample contrast.

A.10 2WikiMultiHopQA Relay versus Final-Answer Effects

We test whether the final answer inherits the relay's intervention effect under single and redundant routing, assessing whether credit can be substituted across scored nodes. Each of the 640 existing episodes contains both outputs; we compute Delete-minus-Control F1 at each node and bootstrap their paired difference by question.

Table A10: Relay and focal credit (Delete minus Control F1). Intervals use the question cluster.

Topology	Relay credit [95% CI]	Focal credit [95% CI]	Focal − relay [95% CI]
Single	-0.0814 $[-0.1223, -0.0448]$	-0.0470 $[-0.0893, -0.0060]$	0.0344 $[0.0094, 0.0656]$
Redundant	-0.3515 $[-0.4263, -0.2775]$	-0.0098 $[-0.0259, 0.0000]$	0.3417 $[0.2675, 0.4171]$

Table A10 shows that the final-answer effect is 0.0344 less negative than the relay effect under single routing and 0.3417 less negative under redundancy; both paired-difference intervals exclude zero. The difference in attenuation between topologies is therefore 0.3073 in the point estimates. The downstream answer does not simply inherit the relay's loss. An untreated route is consistent with attenuation in the redundant graph, but these node comparisons cannot distinguish aggregation, selective evidence use, and repair as mechanisms.

A.11 Full Structural Baseline Family

We test whether the full structural baseline family improves effect prediction over a graph-agnostic score, using the same 320 question–topology paired effects. Fitted predictors use leave-one-question-out evaluation; the deterministic necessity oracle uses these effects without fitting. Graph features are deterministic and computed without outcomes. Reachability records whether an untreated route survives; surviving simple-path count, edge-disjoint path count, and minimum cut describe surviving source–sink connectivity. Shortest-path disconnection records loss of the available path, and untreated bypass count records routes avoiding the treated edge. Every listed feature is one-to-one with topology identity in this two-topology design, so separate feature effects are not identified. Degree and density have no additional variation and supply no separate predictor comparison.

Table A11: Structural baseline family. All metrics are out-of-question leave-one-out errors for the single-effect prediction task.

Predictor	MAE	RMSE	Relation to topology identity
Pooled scalar	0.07321	0.20282	graph-agnostic
Topology label	0.07212	0.20251	reference label
Reachability	0.07212	0.20251	equivalent
Surviving path count	0.07212	0.20251	equivalent
Edge-disjoint paths	0.07212	0.20251	equivalent
Minimum cut	0.07212	0.20251	equivalent
Shortest-path disconnection	0.07212	0.20251	equivalent
Untreated bypass count	0.07212	0.20251	equivalent
Structure + question type	0.08440	0.20116	adds type strata
Structure $\times$ question type	0.08130	0.20039	adds interaction
Deterministic necessity oracle	0.48139	0.70328	binary, wrong scale

Table A11 shows that topology conditioning reduces MAE by 0.00109 and RMSE by 0.00031 relative to the pooled scalar, with every connectivity feature matching the topology-label predictor exactly. Adding question type increases MAE by 0.01228, while including its interaction with structure increases MAE by 0.00918. The identical connectivity errors follow from the feature equivalence and cannot identify separate structural mechanisms. The necessity oracle predicts binary harm when the deleted route is structurally necessary; it provides a qualitative comparator rather than a fitted magnitude model. The natural evidence does not establish a semantic mechanism beyond connectivity.

A.12 Earlier Whole-Message Randomized Interventions

We review earlier whole-message interventions to assess whether their effect magnitudes justify a universal message score. Four predecessor studies use Falcon3 and Tulu3 model families (Falcon Team, 2024; Lambert et al., 2025) on Multi-Hop Question Answering On New And Tail Knowledge (MINTQA), 2WikiMultiHopQA, or the hard subset of the HotpotQA multi-hop question answering dataset (He et al., 2026; Ho et al., 2020; Yang et al., 2018), with a combined reported $n = 3{,}999$. Their intervention units, prompts, models, and outcomes differ from the matched graph studies, so we retain their separate estimates.

Table A12: Earlier randomized whole-message deletion results. Study numbers preserve the reporting order; effects are accuracy percentage-point changes.

Reported study	Deletion effect
Study 1	-6.700
Study 2	-1.300
Study 3	-6.600
Study 4	-4.004
Reported cross-study mean	-3.590

Table A12 summarizes the four deletion losses in their reported order. The largest and smallest losses differ by 5.4 percentage points. Conflict is significantly harmful in three studies, redundancy has mixed effects, and the reported cross-study deletion mean is -3.59 points with heterogeneity $I^2 = 0.922$. This variation prevents the cross-study mean from being interpreted as a universal message effect.

A.13 Earlier Token-Matched Packets and Decision Rules

These earlier studies probe two limits of reusing message-effect estimates: whether communication gains survive token matching and whether attribution-derived rules transfer to multi-hop decisions. Their protocols remain distinct from the matched graph studies.

A.13.1 EXACT-TOKEN MATCHED MESSAGE PACKETS

We examine whether the earlier communication gains persist at fixed prompt length, addressing token count as a possible explanation for intervention effects. The original, deletion-padded, irrelevant, and conflict-padded packets have equal rendered prompt-token counts, allowing comparison at fixed context length. Two Tulu3 protocols on 2WikiMultiHopQA compare communication-enabled and empty-slot conditions.

Table A13: Earlier token-matched evidence. Protocol numbers distinguish the two reported conditions.

Reported protocol	Communication effect (accuracy points)
Protocol 1	-3.33
Protocol 2	1.00

The communication contrast changes by 4.33 percentage points, reversing sign between the protocols (Table A13). A later analysis tests 12 contrasts; none survives Holm correction. Matching tokens therefore controls prompt length without making messages semantically exchangeable or ensuring a consistent accuracy gain.

A.13.2 DECISION RULES AND MULTI-HOP TRANSFER

We examine whether the earlier attribution-derived rules provide reliable multi-hop decisions, testing the downstream use of effect estimates rather than assuming that attribution alone yields a useful policy. Earlier experiments compare attribution-derived rules with verify-all and usefulness-judge policies, using units and outcomes distinct from fixed-execution credit methods such as C3 (Chen et al., 2026).

Table A14: Earlier predefined decision-rule measurements in multi-hop transfer.

Measurement	Reported value
Disagreement	0.333
Regret	0.290
Invalid verifier outputs	$15/108$ (about 13.9%)

Table A14 records the predefined rule's multi-hop transfer measurements. Fifteen invalid outputs among 108 verifier responses amount to about 13.9%, and several predefined support criteria remain unmet. Together with the measured disagreement and regret, these failures prevent the results from establishing a reliable downstream decision policy. They do not evaluate pruning or communication optimization by the two message-credit diagnostics.

B SUPPLEMENTARY METHOD DETAILS

Algorithm 1 specifies how the paired measurements feed the two diagnostics. Its inputs hold the intervention, scored node, model, task distribution, prompting protocol, and scorer fixed, with coverage of every background state receiving positive weight. The procedure operates on completed paired experiments and returns bounds for the supplied estimates, together with the graphs and states attaining their extreme differences. The first output compares graph-level averages; the second compares states within each exposure group. Neither output establishes population-level invariance or predictive accuracy on unmeasured graphs without additional assumptions or evidence.

Algorithm 1 Message-credit portability diagnosis

Require: Paired outcomes for aligned tasks at all states with $Q_G(\mathbf{z}_{-e}) > 0$; graph set $\mathcal{G}$; background weights Q_G; pre-outcome summary ϕ.
Ensure: Scalar minimax error and witness graphs; exposure error bounds and witness states.
Fix intervention, scored node, model, task distribution, prompt protocol, and scorer.
for each evaluated state s **do**
    Compute paired differences $d_i(s)$ and mean $\widehat{C}(s)$ using Equation (4).
end for
Average state credits using Q_G to obtain $\widehat{C}^{Q}_{e,v}(G)$ for every graph.
Find the minimum and maximum graph credits; return c^*_e, $\widehat{\rho}_e/2$, and their graphs.
Group state credits by their common value $a = \phi(s)$.
for each group containing at least two states **do**
    Find its minimum and maximum credits; return $\widehat{D}_\phi(a)/2$ and the two witness states.
end for

C Implementation and Reproducibility

We check which diagnostic calculations the anonymous archive allows readers to verify and which model-backed claims require additional execution inputs. This is a software and evidence-coverage audit, not an additional model experiment. The interface reads Comma-Separated Values (CSV); integrity checks use the Secure Hash Algorithm With A 256-Bit Digest (SHA-256).

Table A15: Contents and verification scope of the anonymous code archive. The toy examples exercise software behavior and are not additional model experiments.

Component	Released content and supported check
Core library	Paired effects, bootstrap helper, graph descriptors, interventions, and finite-design diagnostics.
Aggregate data	36 controlled credits, four maximum collisions, and a natural-task summary with 11 baseline errors.
Protocol templates	Controlled decoding/scoring, natural-study validity fields, and the controlled node prompt.
Validation tools	Summary verifier, CSV analysis interface, two toy inputs, one demonstration, and five assertion tests.
Integrity record	SHA-256 hashes for 27 files, excluding the checksum file itself.

The accompanying code archive separates executable diagnostic logic from model-backed experimental records. Table A15 lists the released components and the checks they support. Its three aggregate data files contain 36 controlled graph-specific credits forming 12 model–task–intervention surfaces, four maximum exposure collisions, and the 160-question natural-task summary with 11 baseline error records. These are existing reported measurements, not new observations. They allow the range, half-range, half-diameter, and natural graph-contrast arithmetic to be checked directly; they do not contain the paired outcomes required to recompute confidence intervals or refit the baseline predictors.

Message events store original and observed payloads separately. Delete writes `[DELETED]` and empties observed atom identifiers; Conflict preserves the event metadata and checks that the substituted text matches the assignment's stored SHA-256 hash. The controlled prompt requests a local answer, selected fact identifiers, and confidence, and requires abstention when evidence is missing. Configuration fixes temperature to zero and limits new tokens to 64. These mechanisms expose the intervention actually delivered to a node without interpreting its subsequent output.

Once state effects and exposure labels are available, the range and exposure-group scans require linear work in the number of states, apart from ordering labels for display; this excludes model execution and graph-feature computation. The scalar helper returns the minimum, maximum, range, midpoint, and half-range. Recovering the corresponding graph identities requires an additional extrema lookup in the supplied surface, as specified by the diagnostic procedure. The exposure helper

directly returns minimum- and maximum-effect state identities for each observed summary value. The released group-mean helper excludes every record sharing the held-out unit identifier, preventing a question's other graph from entering its training group; groups with no remaining training examples receive no prediction.

The aggregate verifier, toy demonstration, CSV interface, and all five unchanged assertion-test functions pass, and all 27 listed file hashes match. These checks validate the released calculations and software behavior. Model weights, provider backends, generated traces, benchmark question records, and full experiment drivers are excluded. The package consequently supports aggregate verification and reuse of the analysis components, while independent model-backed replication additionally requires those experiment inputs and execution infrastructure.