DASH-AGENT: RELIABLE DISSENT PRESERVATION FOR LLM MULTI-AGENT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM)-based multi-agent systems benefit from diverse reasoning and evidence, yet coordination can erase the information that makes this diversity useful. When several agents share an error, consensus may suppress a minority claim carrying the evidence needed to correct the final answer. Preserving every disagreement instead propagates noise and consumes verification budget. Rather than treating consensus or sparse communication as the endpoint, we study *reliable dissent preservation*: retaining evidence-bearing minority claims while filtering unsupported deviations under bounded communication. DASH-Agent implements this objective with a claim-level discovery–validation–verification pipeline. A task-conditioned interaction graph exposes candidate dissent, Dissent Signal Triage (DST) ranks candidates by novelty, task utility, reliability, and cost, and sparse local groups verify selected claims before synthesis. Extensive experiments across reasoning, multi-hop question answering, and tool-use benchmarks show improved task quality with lower communication, while claim-level analyses show stronger useful-dissent retention and incorrect-dissent suppression.

1 INTRODUCTION

Large language model (LLM)-based agents increasingly solve complex tasks through debate, role specialization, retrieval, and tool use (Du et al., 2024; Wu et al., 2024; Zhang et al., 2025). Their value partly comes from exposing complementary information that a single agent may miss. The same diversity, however, creates a coordination problem: several agents can converge on an incorrect assumption while a minority agent holds the passage, derivation, or tool observation needed to correct it. If aggregation compresses these responses into early consensus, the corrective evidence can disappear. We call this failure *dissent collapse*. Prior work has documented majority-convergence failures and shown that communication topology changes how both correct and erroneous information propagate through multi-agent systems (Estornell & Liu, 2024; Shen et al., 2025).

Simply preserving every disagreement does not solve the problem. A message may contain overlooked evidence, but it may also be irrelevant, hallucinated, or internally inconsistent. Recent systems improve debate, confidence-weighted consensus, sycophancy mitigation, workflow design, and communication sparsity (Chen et al., 2024; Pitre et al., 2025; Verma et al., 2026; Zhang et al., 2024; Yun et al., 2026). Yet final-answer accuracy does not show whether supported minority evidence survived coordination, and lower token use can reflect the removal of either noise or useful dissent. The central question is therefore not only *who should communicate*, but *which dissenting claims deserve verification and preservation*. Figure 1 summarizes this tension between majority convergence, indiscriminate disagreement, and focused checking.

DASH-Agent answers this question through three claim-level decisions. As shown in Figure 2, graph variation controls which claims become visible without assigning truth, Dissent Signal Triage (DST) prioritizes exposed claims, and compatibility-based local groups adjudicate selected claims before synthesis. Because a response can mix useful evidence with unsupported inference, claim-level coordination routes claims independently rather than applying one trust judgment to the entire speaker.

The main contributions are:

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.

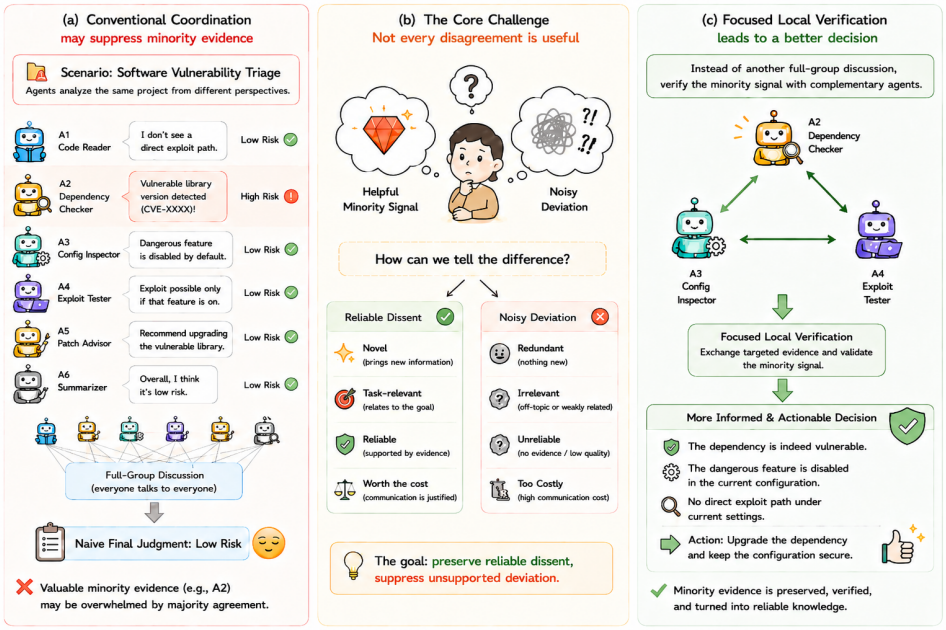

Figure 1: Motivation for reliable dissent preservation. Majority agreement can suppress corrective minority evidence, while novelty alone does not establish correctness. Selected claims are therefore verified before synthesis.

- **Distinct coordination objective.** Reliable dissent preservation is formulated as a budgeted claim-level objective that targets the survival of corrective minority evidence rather than consensus or communication sparsity alone.
- **Decision decomposition.** DASH-Agent separates exposure, prioritization, and adjudication into non-interchangeable stages, allowing claims from the same speaker to receive different routing and verification outcomes.
- **Mechanism-oriented evaluation.** Calibration, candidate recall, Minority-Correct Retention (MCR), Useful-Dissent Survival (UDS), and Incorrect-Dissent Suppression (IDS) trace whether corrective evidence survives coordination.
- **Broad empirical comparison.** Across reasoning, multi-hop question answering, and tool-use tasks, DASH-Agent consistently improves task-level performance over recent multi-agent coordination baselines while using less communication.

2 Related Work

Our work connects three lines of research: consensus and debate in LLM-based multi-agent systems, adaptive communication structures, and graph-based reliability modeling.

Consensus, debate, and dissent. Multi-agent debate exchanges rationales before aggregation, while ReConcile adds iterative discussion with confidence-weighted consensus (Du et al., 2024; Chen et al., 2024). CONSENSAGENT targets sycophantic convergence, and SELENE selectively initiates debate while weighting evidence during aggregation (Pitre et al., 2025; Verma et al., 2026). Estornell and Liu further analyze majority-convergence failures and interventions in multi-LLM debate (Estornell & Liu, 2024). Their primary objectives concern answer quality or the consensus process. DASH-Agent instead makes the survival of evidence-supported minority claims an explicit coordination objective and evaluates that survival directly.

Adaptive communication and structured coordination. AgentPrune, AFlow, and Graph-of-Agents adapt communication pipelines, workflows, or interaction graphs to improve coordination

or reduce unnecessary exchange (Zhang et al., 2024; 2025; Yun et al., 2026). Communication topology can itself alter the propagation of correct and incorrect information (Shen et al., 2025). These approaches primarily decide who communicates, when, or over which topology. DASH-Agent instead uses topology to expose claim candidates before separate evidence-based prioritization. Higher-order grouping, also explored by HYGMA and EvoHyper (Liu & Li, 2025; Zhang et al., 2026), is used only downstream for focused adjudication.

Graph signals and reliability. Graph signal processing separates locally smooth and high-variation components on irregular structures (Shuman et al., 2013), while calibration connects predictive scores to empirical correctness (Guo et al., 2017). Related graph-learning work likewise combines structural information with reliability-aware guidance. GNN-LS, for example, uses information enhancement and reliability-guided supervision to address label scarcity on graphs (Yang et al., 2026). DASH-Agent does not interpret graph variation as reliability. It uses variation only to expose potentially informative deviations, then lets task evidence and calibrated reliability determine which claims deserve verification.

3 Reliable Dissent Preservation Framework

The framework treats reliable dissent preservation as three distinct claim-level decisions rather than one agent-level routing score: exposure, verification priority, and evidence-based adjudication. Section 3.1 defines the objective, Sections 3.2–3.3 implement discovery and triage, and Section 3.4 verifies selected claims.

3.1 Problem Formulation

For a query q, let N agents produce initial answers a_i, messages m_i, and optional evidence e_i. At communication round t, agent i has state $x_i^t = f(q, r_i, m_i^t, e_i^t, \kappa_i^t, \bar{R}_i^t)$, where r_i is its role, κ_i^t is reported confidence, $\bar{R}_i^t$ is prior reliability, and $f(\cdot)$ constructs the state representation. The initial majority answer is a_{maj}, and the answer-level dissent set is $\mathcal{D} = \{i : a_i \neq a_{\mathrm{maj}}\}$. Claim-level dissent is broader because agents can agree on the final answer while relying on conflicting evidence. A speaker-level gate ties all claims from one response to one routing decision, whereas claim-level coordination can retain a supported claim and reject an unsupported claim from the same response.

For evaluation, a dissenting claim is *reliable* when independent task evidence supports it and the claim can materially contribute to the reference resolution. This post-hoc label is never exposed to the test-time policy. Let $\mathcal{C}^t$ be the candidate claims available at round t. For claim $c \in \mathcal{C}^t$, $\hat{r}(c)$ denotes estimated reliability, $u(c)$ estimated task utility, and $b(c)$ verification cost. Under per-round budget B_t, the target selection problem is

$$\mathcal{V}^{t*} = \arg\max_{\mathcal{V} \subseteq \mathcal{C}^t} \sum_{c \in \mathcal{V}} \hat{r}(c) u(c) \quad \text{s.t.} \quad \sum_{c \in \mathcal{V}} b(c) \leq B_t. \tag{1}$$

Equation 1 exposes three non-interchangeable decisions: discovery sets the feasible claim set $\mathcal{C}^t$, triage allocates verification capacity within it, and verification determines support before synthesis. A missed claim cannot be recovered downstream, while high priority alone cannot establish evidential support. DASH-Agent therefore separates exposure, priority, and acceptance instead of collapsing them into one routing score. The graph neighborhood size, claim-admission threshold, DST weights, routing cap, group threshold, and stopping rule are selected on held-out validation data and frozen at test time. Held-out evidence labels calibrate reliability, while test labels are never used for routing or parameter selection. Full generation and routing specifications are provided in Appendix A.

3.2 Candidate Dissent Discovery

The first stage determines which differences should be exposed to later verification. At round t, symmetric adjacency A^t averages four fixed $[0, 1]$ cues—task relevance, role complementarity, answer conflict, and evidence compatibility—then retains the union of each agent's top-K_s neighbors. No learned edge predictor or test label enters graph construction. Let D^t be the degree matrix, I the identity matrix, and $L^t = I - (D^t)^{-1/2} A^t (D^t)^{-1/2}$ the normalized graph Laplacian (Shuman

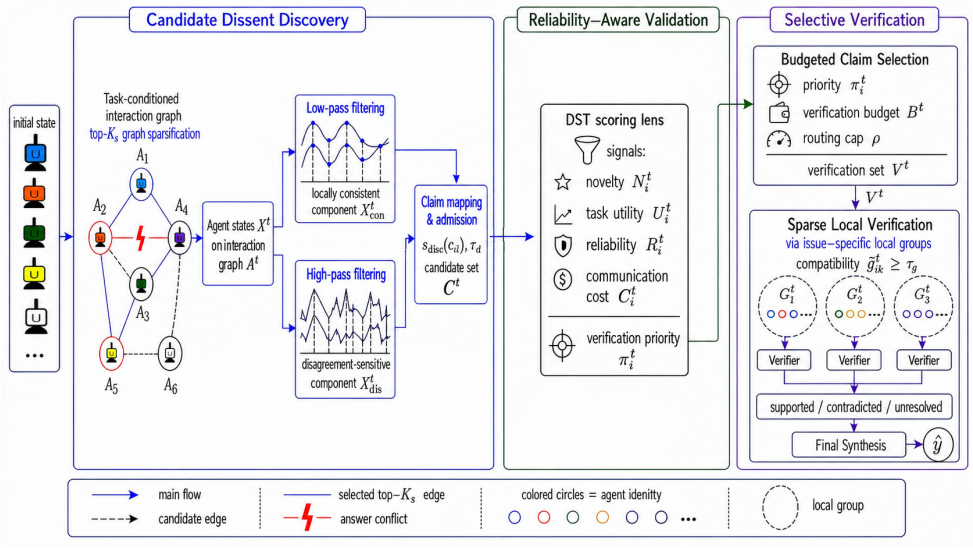

Figure 2: DASH-Agent pipeline. Candidate Dissent Discovery exposes disagreement-sensitive claims, DST prioritizes them by novelty, task utility, reliability, and cost, and Selective Verification routes only selected claims to issue-specific local groups before adjudication and synthesis. Formal notation is defined in Sections 3.2–3.4.

et al., 2013). For stacked state matrix X^t, complementary filters produce locally consistent and disagreement-sensitive components:

$$X^t_{\text{con}} = (I - \tfrac{1}{2}L^t)X^t, \qquad X^t_{\text{dis}} = \tfrac{1}{2}L^t X^t. \tag{2}$$

Here X^t_{con} emphasizes information smooth over the interaction graph and X^t_{dis} local variation. Proposers emit structured claim–evidence records $c_{i\ell}$. Discovery assigns each record a score $d^t_{i\ell}$ that combines source-level graph variation with claim-level answer-implication conflict and evidence divergence from the source agent's neighborhood. Only records above a validation-selected threshold τ_d enter $\mathcal{C}^t$. Thus two claims from the same agent can receive different admission decisions even though they share the source variation signal. Graph variation remains evidence of disagreement rather than correctness, and later DST separately evaluates utility and reliability. Exact scoring and admission rules are given in Appendix A. Section 4.4 evaluates the resulting exposure with candidate recall and useful precision.

3.3 Reliability-aware Validation

Once candidate claims are exposed, Dissent Signal Triage (DST) converts four claim-level signals into a verification priority. Novelty N^t_i derives from graph variation, utility U^t_i measures whether resolving the claim can change the task state, and cost C^t_i estimates additional tokens, latency, and tool calls. Reliability R^t_i averages the available message–evidence consistency, executed-tool validity, reported confidence, and held-out source-reliability cues, then applies a one-dimensional calibration map fitted only on held-out evidence labels and frozen at test time. N^t_i, U^t_i, and C^t_i are scaled to $[0, 1]$ within each round, while $R^t_i \in [0, 1]$ retains its cross-query interpretation.

The four signals form $\mathbf{z}^t_i = [N^t_i, U^t_i, R^t_i, -C^t_i]^\top$. A fixed weight vector $\mathbf{w} \in \mathbb{R}^4$, chosen with the other coordinator hyperparameters on the validation split, produces triage score S^t_i. A softmax over the candidate set converts these scores into relative verification priority π^t_i:

$$S^t_i = \mathbf{w}^\top \mathbf{z}^t_i, \qquad \pi^t_i = \frac{\exp(S^t_i)}{\sum_{j \in \mathcal{C}^t} \exp(S^t_j)}. \tag{3}$$

Priority π^t_i ranks claims for verification, while the later verifier determines whether a claim is supported. The softmax is recomputed over the current candidate set, so π^t_i expresses within-round priority rather than a probability of correctness. The four signals therefore answer different questions.

N_i^t asks whether the claim departs from its local context, U_i^t asks whether resolving it can change the current task state, R_i^t asks whether its evidence is likely to withstand checking, and C_i^t asks how expensive that check will be. A highly novel claim can consequently receive low priority when it lacks support, whereas a moderately novel claim can move upward when it is both decisive and well grounded. This separation prevents disagreement magnitude from acting as a surrogate for truth.

Evidence takes the form appropriate to the task: an independent derivation for mathematical reasoning, supporting passages for multi-hop question answering, or executed application programming interface (API) outputs and schema checks for tool use. Because R_i^t is computed before selective verification, its calibration can be evaluated against evidence-based correctness before the later adjudication step.

3.4 Selective Verification

The final stage adjudicates high-priority claims without reopening a full-group discussion. Let b_i^t be the cost of verifying candidate i and B_t the per-round verification budget. DASH-Agent selects the highest-priority subset subject to this budget. For selected issue k, a local group is formed from agents whose compatibility score $\widetilde{g}_{ik}^t$ exceeds threshold τ_g:

$$\mathcal{V}^t = \arg\max_{\mathcal{S}\subseteq\mathcal{C}^t} \sum_{i\in\mathcal{S}} \pi_i^t \quad \text{s.t.} \quad \sum_{i\in\mathcal{S}} b_i^t \le B_t, \qquad \mathcal{G}_k^t = \{i : \widetilde{g}_{ik}^t \ge \tau_g\}. \tag{4}$$

Here $\mathcal{V}^t$ is the verification set, $\mathcal{G}_k^t$ the local group for issue k, and τ_g the compatibility threshold. Compatibility favors agents with task-relevant evidence that is complementary to the source of the disputed claim. The reported policy routes at most half of the discovered candidate claims. The verifier inspects the claim, its source, supporting evidence, and relevant counterevidence, then returns *supported*, *contradicted*, or *unresolved*. Final synthesis retains supported claims, excludes contradicted claims, and marks unresolved claims as uncertain.

Verification groups are issue-specific adjudicators rather than persistent agent teams or a general collaboration topology. An agent can therefore be useful for one disputed claim and irrelevant to another, preserving role and evidence complementarity where it matters. Unrelated agents stay outside the exchange, so additional communication is spent on adjudicating a concrete claim rather than reopening the entire discussion.

Claim-specific verification allows different statements from the same agent to receive different outcomes. Coordination stops after at most three rounds or when post-verification consensus confidence no longer improves. For d-dimensional states and graph edge set E, pairwise graph construction costs $O(N^2 d)$ and sparse filtering costs $O(|E|d)$. Verification then scales with the routed claim set rather than another full-agent broadcast, which motivates the communication analysis in Section 4.6.

4 Experiments

Because the method separates exposure, prioritization, and adjudication, the experiments test these decisions along the same evidence path. We first establish final-task quality, evaluate the reliability signal, trace candidate dissent into claim-level preservation, perturb each stage, and finally measure the communication required to obtain these gains.

4.1 Experimental Setup

To isolate coordination effects, all primary comparisons use the same Qwen2.5-7B-Instruct backbone and task metrics. The benchmark suite contains GSM8K and MATH for mathematical reasoning, HotpotQA for multi-hop question answering, and API-Bank and ToolBench for tool use (Cobbe et al., 2021; Hendrycks et al., 2021; Yang et al., 2018; Li et al., 2023; Qin et al., 2024). GSM8K and MATH use accuracy, HotpotQA uses answer F1 score, and API-Bank and ToolBench use task success. Each run starts with eight agents—six proposers and two verifiers—and allows at most three coordination rounds. Generation uses temperature 0.7, top-p 0.9, and a maximum of 2048 tokens. Results are reported over five random seeds.

The comparison set includes Debate, Graph-of-Agents, AFlow, AgentPrune, ReConcile, CONSENSAGENT, and SELENE (Du et al., 2024; Yun et al., 2026; Zhang et al., 2025; 2024; Chen et al., 2024;

Table 1: Five-benchmark comparison (percent, mean $\pm$ standard deviation over five seeds). Best means are bolded and runner-up means are underlined.

Method	GSM8K	MATH	HotpotQA	API-Bank	ToolBench
Debate	77.4 ± 1.1	44.3 ± 1.2	67.6 ± 1.0	57.9 ± 1.1	52.8 ± 1.0
Graph-of-Agents	79.0 ± 0.8	45.1 ± 0.9	68.9 ± 0.8	59.1 ± 0.9	54.0 ± 0.8
AFlow	78.7 ± 1.0	44.9 ± 1.1	68.6 ± 0.9	58.8 ± 0.7	53.7 ± 0.9
AgentPrune	78.1 ± 1.1	44.5 ± 1.0	68.0 ± 1.1	58.2 ± 0.9	53.2 ± 1.2
ReConcile	79.3 ± 0.7	45.5 ± 0.8	69.4 ± 0.7	59.4 ± 0.8	54.4 ± 0.8
CONSENSAGENT	79.7 ± 0.6	45.9 ± 0.7	69.8 ± 0.7	59.8 ± 0.7	54.9 ± 0.7
SELENE	$\underline{80.1 \pm 0.6}$	$\underline{46.3 \pm 0.7}$	$\underline{70.2 \pm 0.6}$	$\underline{60.2 \pm 0.6}$	$\underline{55.3 \pm 0.8}$
DASH-Agent	$\mathbf{83.4 \pm 0.5}$	$\mathbf{47.6 \pm 0.6}$	$\mathbf{72.9 \pm 0.5}$	$\mathbf{63.0 \pm 0.6}$	$\mathbf{57.2 \pm 0.5}$

Table 2: Observed correctness across pre-verification reliability-score intervals (percent, mean $\pm$ standard deviation over five seeds).

Lower-score intervals		Higher-score intervals	
Score interval	**Correctness**	**Score interval**	**Correctness**
$[0.0, 0.1)$	29.8 ± 3.4	$[0.5, 0.6)$	71.7 ± 1.8
$[0.1, 0.2)$	37.6 ± 2.7	$[0.6, 0.7)$	78.9 ± 1.6
$[0.2, 0.3)$	46.9 ± 3.0	$[0.7, 0.8)$	85.2 ± 1.7
$[0.3, 0.4)$	55.3 ± 2.1	$[0.8, 0.9)$	90.8 ± 1.1
$[0.4, 0.5)$	63.8 ± 2.5	$[0.9, 1.0]$	94.6 ± 0.9

Pitre et al., 2025; Verma et al., 2026). All methods use the shared backbone. The two routing controls also reuse the same initial responses and verification-slot budget. Random routing samples uniformly from DASH-Agent's graph-exposed candidate set, whereas reliability-only routing bypasses graph discovery and ranks explicit answer-dissent claim records by calibrated reliability alone. The mechanism audit uses a preselected 1,200-instance stratified sample with two independent annotators, blinded routing outcomes, and third-annotator adjudication. Full counts, agreement statistics, and accounting rules are provided in the Appendix.

4.2 Main Benchmark Results

Table 1 first tests final-task quality. DASH-Agent achieves the highest mean on all five tasks, exceeding SELENE by 3.3, 1.3, 2.7, 2.8, and 1.9 points on GSM8K, MATH, HotpotQA, API-Bank, and ToolBench, respectively.

The gains span reasoning, multi-hop question answering, and tool use. Because final-task scores do not reveal whether corrective minority evidence survives coordination, the remaining experiments trace the internal evidence path.

4.3 Reliability Calibration

Before evaluating preservation, we test whether the pre-verification reliability score used by DST tracks empirical correctness. Expected Calibration Error (ECE) and the Brier score are standard calibration measures (Guo et al., 2017). Table 2 groups claims into ten reliability-score intervals and reports the fraction supported by independent task evidence. Observed correctness rises monotonically from $29.8 \pm 3.4\%$ in $[0.0, 0.1)$ to $94.6 \pm 0.9\%$ in $[0.9, 1.0]$.

At claim level, DST yields ECE 0.053 ± 0.007 and Brier score 0.087 ± 0.009, computed from individual scores rather than interval midpoints. The lower-score bins remain conservative, so these values are interpreted as aggregate calibration under the empirical candidate distribution rather than evidence that every interval is perfectly calibrated. Per-bin support and mean score are reported in the Appendix for direct verification.

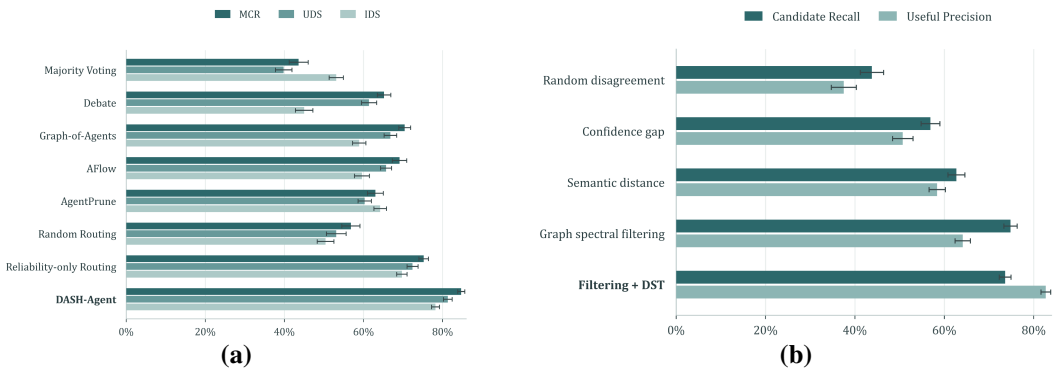

Figure 3: Claim-level mechanism analysis (percent, mean $\pm$ standard deviation). (a) Preservation and suppression by routing method. (b) Candidate recall and useful precision across discovery and triage rules.

4.4 Reliable Dissent Preservation and Discovery

We next test whether supported minority evidence survives while unsupported dissent is filtered. With the initial majority fixed before coordination, annotated claims are judged against independent task evidence. Minority-Correct Retention (MCR) measures eligible wrong-majority instances retaining at least one corrective minority claim, Useful-Dissent Survival (UDS) measures the fraction of useful dissent retained, and Incorrect-Dissent Suppression (IDS) measures the fraction of incorrect dissent excluded. Preservation requires availability to final synthesis, not mere appearance in the transcript.

Figure 3(a) shows that DASH-Agent reaches $84.7 \pm 0.9\%$ MCR, $81.3 \pm 1.1\%$ UDS, and $78.2 \pm 1.0\%$ IDS. Reliability-only routing reaches $75.2 \pm 1.2\%$, $72.4 \pm 1.4\%$, and $69.7 \pm 1.3\%$, leaving gaps of 9.5, 8.9, and 8.5 points. It can rank explicit answer-dissent claims by reliability but lacks graph-conditioned exposure of evidence-level conflicts that need not change a speaker's final answer. Debate shows the complementary failure mode: stronger retention than majority voting but weaker suppression. The pattern therefore favors exposure followed by selective filtering rather than either mechanism alone.

The three rates are complementary: MCR is instance-level, UDS measures broader useful-claim survival, and IDS prevents aggressive rejection from appearing strong merely by suppressing dissent.

Figure 3(b) isolates discovery from validation. Graph spectral filtering reaches $74.8 \pm 1.5\%$ candidate recall and $64.1 \pm 1.7\%$ useful precision. Adding DST yields $73.6 \pm 1.3\%$ recall and $82.7 \pm 1.1\%$ precision, an 18.6-point precision gain for a 1.2-point recall decrease. This matches the intended division of labor: discovery broadens exposure while DST concentrates verification on better-supported candidates.

4.5 Ablation and Mechanism Gaps

The preceding analysis shows how claims move through the pipeline. We next test whether each stage is necessary for task-level gains and whether simpler routing policies reproduce this mechanism.

Removing discovery, DST, or selective verification lowers the three-task average by 1.9, 3.4, and 2.5 points, while reliability-only and random routing fall by 4.4 and 6.0 points. These interventions are sequential rather than additive: discovery defines the feasible claim set, DST allocates verification capacity, and verification adjudicates routed claims.

Figure 4(b) gives the corresponding claim-level view. Reliability-only routing remains closest but trails DASH-Agent by 9.5 points in MCR, 8.9 in UDS, and 8.5 in IDS. The fourth column is only a descriptive mean because the three metrics use different populations. Together with Figure 3, these ablations tie task gains to the evidence path rather than sparsity or speaker trust.



Figure 4: Ablation and mechanism gaps (percentage points). (a) Task-score change from full DASH-Agent on GSM8K, MATH, HotpotQA, and their unweighted mean. (b) MCR, UDS, and IDS gaps from DASH-Agent, with an unweighted descriptive average shown only as a compact summary.

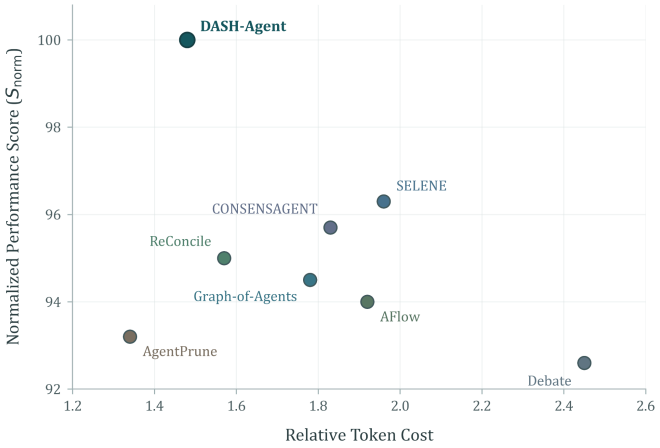

Figure 5: Communication–performance trade-off, where lower cost and higher S_{norm} are preferred.

4.6 Communication Efficiency

Given that focused verification is intended to avoid unnecessary full-group interaction, an important question is whether its quality gains are obtained without substantially increasing communication. To compare heterogeneous task metrics on a common scale, S_{norm} first normalizes each method's score by the best value on the same benchmark and then averages the five normalized scores. Exchanged-token cost includes initial generation, routed evidence, verification, judging, synthesis, and tool exchanges where applicable. Figure 5 reports the resulting cost–quality trade-off.

DASH-Agent attains the highest normalized performance, $S_{\text{norm}} = 100.0$, at a relative token cost of 1.5, compared with Debate at 2.5. This corresponds to roughly $0.6\times$ Debate's communication cost, or a reduction of about 40%. SELENE reaches $S_{\text{norm}} = 96.3$ at cost 2.0, while Graph-of-Agents and AFlow obtain 94.5 and 94.0 at costs 1.8 and 1.9, respectively. AgentPrune communicates less at 1.3, but its lower normalized performance shows that reducing interaction alone is insufficient.

This pattern complements the mechanism results in Sections 4.4 and 4.5. Candidate discovery enlarges the set of potentially corrective claims, whereas only a subset is subsequently routed for checking. DST and the hard verification budget therefore concentrate communication on claims that are both potentially useful and worth adjudicating, rather than repeatedly reopening the full discussion. The gain in Figure 5 should consequently be interpreted as a joint quality–cost outcome of selective coordination rather than as evidence that fewer tokens are intrinsically better. Original task metrics remain reported in Table 1, while S_{norm} is used only for this cross-task display.

4.7 Evidence Chain and Generalization

The experiments collectively follow the proposed decision decomposition. Table 1 establishes final-task quality, while Figure 3 separates candidate exposure from useful filtering and Figure 4 tests the contribution of discovery, DST, and selective verification. Figure 5 further shows the communication required to obtain these gains. This evidence chain helps distinguish improvements caused by different coordination mechanisms rather than relying on final-task accuracy alone.

The paired controls further separate exposure from prioritization. Random routing uses DASH-Agent's graph-exposed candidate set under the same verification-slot budget, isolating the value of structured prioritization once exposure is fixed. Reliability-only routing instead shares the initial responses and verification capacity but constructs candidates from explicit answer dissent without graph-conditioned discovery. The two controls therefore probe different stages of the pipeline rather than interchangeable routing variants.

The pattern also extends beyond the primary Qwen2.5-7B-Instruct setting. With Llama-3.1-8B-Instruct, DASH-Agent obtains 80.2 ± 0.7 on GSM8K, while CommonsenseQA and StrategyQA reach 78.6% and 74.4%, exceeding the strongest listed alternatives by 1.3 and 1.5 points. Additional transfer results are reported in Appendix D. Calibration is likewise interpreted as support for prioritization rather than a truth criterion: lower-score intervals remain conservative, and aggregate ECE reflects the empirical candidate distribution. DST therefore combines reliability with novelty, task utility, and checking cost before verification determines the final evidential disposition.

4.8 Discussion and Failure Modes

The combined evidence supports the central modeling choice that exposure, verification priority, and evidential acceptance should remain separate claim-level decisions. A response can contain both useful evidence and unsupported inference, and agents can agree on an answer while relying on incompatible evidence. Speaker-level trust or answer-level disagreement alone therefore cannot express all cases that matter for corrective coordination.

This decomposition also yields a direct failure taxonomy. A *discovery failure* occurs when a useful claim does not enter the candidate set because its source is not sufficiently exposed or its claim-level conflict cues remain weak. A *validation failure* places an exposed claim too low under the verification budget. A *verification failure* occurs when the selected evidence is incomplete or insufficient to resolve the dispute. Candidate recall primarily diagnoses discovery, useful precision and calibration diagnose prioritization, while strong upstream precision combined with weak MCR or UDS points to the verification or synthesis path. The same final-answer error can therefore arise from different coordination failures and require different interventions.

The formulation is most useful when agents can expose checkable intermediate evidence such as derivations, supporting passages, or tool observations. It cannot recover evidence that no agent observes, and its benefits may diminish when many useful claims compete for a small verification budget or when no stable evidential criterion is available. These limits motivate future work on adaptive discovery, verification budgets, and local-group construction while preserving the claim-level separation that makes the coordination process auditable.

5 Conclusion

DASH-Agent reframes LLM multi-agent coordination around reliable dissent preservation, separating claim exposure, evidence-aware prioritization, and selective adjudication so minority evidence can remain actionable without treating disagreement as truth. This claim-level decomposition keeps routing priority distinct from correctness and localizes failures to discovery, ranking, or verification. Across reasoning, multi-hop question answering, and tool use, the evidence path supports stronger task quality and useful-dissent retention under bounded communication. The framework remains limited by unobserved evidence, distribution-dependent calibration, and finite verification capacity. Reliable coordination should therefore preserve the evidence needed to correct consensus rather than optimize consensus alone. Future work will study heterogeneous agents, distribution-shift calibration, auditable learned discovery, and adaptive graph and verification-budget design.

AI USAGE DISCLOSURE

Generative AI tools were used for manuscript editing, language refinement, and LaTeX preparation. All scientific contributions, methodology, experimental analyses, and conclusions were developed and verified by the authors.

ETHICS STATEMENT

This work does not involve human-subject studies or the collection of personally identifiable information. Experiments use publicly available benchmarks and model-generated outputs. The authors adhere to the ICLR Code of Ethics throughout the research and submission process.

REPRODUCIBILITY STATEMENT

Reproducibility details are provided throughout the paper and appendix. The appendix includes the complete DASH-Agent procedure, fixed experimental configuration, theoretical derivations, claim-annotation and calibration protocols, extended numerical results, communication accounting, and prompt specifications. These materials are intended to support independent reproduction of the reported experiments.

REFERENCES

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://aclanthology.org/2024.acl-long.381/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*, 2024.

Andrew Estornell and Yang Liu. Multi-LLM debate: Framework, principles, and interventions. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28938–28964, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

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.

Minghao Li, Yijun Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. API-Bank: A comprehensive benchmark for tool-augmented LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3102–3116, 2023.

Chiqiang Liu and Dazi Li. HYGMA: Hypergraph coordination networks with dynamic grouping for multi-agent reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38767–38788, 2025.

Priya Pitre, Naren Ramakrishnan, and Xuan Wang. CONSENSAGENT: Towards efficient and effective consensus in multi-agent LLM interactions through sycophancy mitigation. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025. URL https://aclanthology.org/2025.findings-acl.1141/.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations*, 2024.

Xu Shen, Yixin Liu, Yiwei Dai, Yili Wang, Rui Miao, Yue Tan, Shirui Pan, and Xin Wang. Understanding the information propagation effects of communication topologies in LLM-based multi-agent systems. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12347–12361, 2025. doi: 10.18653/v1/2025.emnlp-main.623.

David I. Shuman, Sunil K. Narang, Pascal Frossard, Antonio Ortega, and Pierre Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, 2013.

Akshay Verma, Swapnil Gupta, Deepak Gupta, Prateek Sircar, and Siddharth Pillai. SELENE: Selective and evidence-weighted LLM debating for efficient and reliable reasoning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 95–104, 2026. URL https://aclanthology.org/2026.eacl-industry.7/.

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

Haoran Yang, Junli Wang, Rui Duan, Xin Guo, and Chungang Yan. GNN-LS: An information-enhanced method for label scarcity on graphs. *ACM Transactions on Knowledge Discovery from Data*, 20(8):1–24, 2026. doi: 10.1145/3837756.

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, 2018.

Sukwon Yun, Jie Peng, Pingzhi Li, Wendong Fan, Jie Chen, James Zou, Guohao Li, and Tianlong Chen. Graph-of-agents: A graph-based framework for multi-agent LLM collaboration. *arXiv preprint arXiv:2604.17148*, 2026.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. *arXiv preprint arXiv:2410.02506*, 2024.

Heng Zhang, Yihao Zhong, Lubin Gan, Zhihe Chen, Jiajun Wu, Yuling Shi, Xiaodong Gu, Hao Zhang, Haochen You, and Jin Huang. EvoHyper: Evolving hypergraph topologies for unified collaboration in multi-agent communication. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 25130–25143, 2026.

Jiayi Zhang, Jiaming Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. AFlow: Automating agentic workflow generation. In *International Conference on Learning Representations*, 2025.

A Full Algorithm and Reproducibility Details

The appendix complements the main paper without changing the test-time policy described in Section 3. We first give an end-to-end specification of DASH-Agent, then formalize properties of the discovery–validation–verification decomposition, and finally provide the evaluation and accounting details underlying the reported results. All primary experiments use the configuration frozen on held-out validation data before test evaluation.

A.1 End-to-end DASH-Agent procedure

Let q be the query, $\mathcal{A} = \{1, \ldots, N\}$ the agent set, B_t the round-t verification-slot budget, K_s the graph neighborhood size, w the Dissent Signal Triage (DST) weight vector, τ_g the local-group compatibility threshold, and T the maximum number of coordination rounds. Algorithm 1 makes the decision decomposition explicit. Graph variation exposes candidate claims, DST orders exposed claims, and verification assigns evidential disposition.

Algorithm 1 DASH-Agent for Reliable Dissent Preservation

Require: Query q, agents $\mathcal{A}$, verification-slot budget B_t, neighborhood size K_s, claim threshold τ_d, DST weights w, compatibility threshold τ_g, maximum rounds T
Ensure: Final answer $\hat{y}$ and verified claim record $\mathcal{R}$

```
Generate each agent's initial answer, structured claims, evidence links, confidence, and role-conditioned message
Construct X^0 and initial majority answer a_maj
Initialize R ← ∅ and previous synthesis confidence γ_{-1} ← −∞
for t = 0, ..., T − 1 do
    Construct the symmetric task-conditioned graph A^t and retain the union of top-K_s neighbors
    Compute L^t and filtered states X^t_con = (I − (1/2)L^t)X^t and X^t_dis = (1/2)L^t X^t
    Score structured claims with graph-conditioned claim divergence and admit those above τ_d into C^t
    for all c_i ∈ C^t do
        Compute novelty N^t_i, utility U^t_i, reliability R^t_i, and communication cost C^t_i
        Compute S^t_i = w^⊤ [N^t_i, U^t_i, R^t_i, −C^t_i]^⊤ and priority π^t_i using Equation 3
    end for
    K_t ← min{B_t, ⌊0.5|C^t|⌋} and V^t ← the K_t highest-priority claims
    for all c_i ∈ V^t do
        Form an issue-specific local group from agents with compatibility at least τ_g
        Verify c_i against source evidence and relevant counterevidence
        Assign supported, contradicted, or unresolved, then update R
    end for
    Produce an interim synthesis and post-verification consensus confidence γ_t
    if t > 0 and γ_t ≤ γ_{t−1} then
        break
    end if
end for
Produce ŷ from the initial responses and R, retaining support and marking unresolved claims as uncertain
```

The reported implementation expresses the hard round budget in verification slots, so each routed claim consumes one slot after the claim packet is length capped. The continuous communication estimate C_i^t remains a DST feature and is not reused as the hard slot cost. Under this implementation, Step 11 is the exact solution of the budgeted selection rule in the main paper. The more general heterogeneous-cost form is analyzed in Section B.4. The procedure is intentionally claim-level. One response can therefore contribute a supported claim and a contradicted claim without forcing both to share a speaker-level trust decision.

A.2 Fixed configuration

All primary benchmark, mechanism, calibration, ablation, and communication experiments use Qwen2.5-7B-Instruct. Each task starts with eight agents, consisting of six proposers and two verifiers. The maximum number of coordination rounds is three. Generation uses temperature 0.7, top-p 0.9, and a maximum generation length of 2048 tokens. Candidate discovery uses $K_s = 4$ and

claim-admission threshold $\tau_d = 0.45$. The fixed DST weight vector is $(0.30, 0.25, 0.35, 0.10)$ for novelty, task utility, reliability, and communication cost, with cost entering negatively. At most half of the discovered candidate claims are routed in a round. These values are selected on held-out validation data and frozen for all reported test runs.

Table A.1: Coordinator configuration used in the reported experiments.

Item	**Setting**
Agent and generation configuration	
Primary backbone	Qwen2.5-7B-Instruct
Agent composition	6 proposers + 2 verifiers
Maximum coordination rounds	3
Temperature / top-p	0.7 / 0.9
Maximum generation length	2048 tokens
Graph, triage, and routing configuration	
Graph neighborhood size K_s	4
Claim-admission threshold τ_d	0.45
DST weights (N, U, R, C)	$(0.30, 0.25, 0.35, 0.10)$ with cost entering negatively
Hard verification budget	unit verification slots
Candidate-routing cap	at most 50% of discovered claims per round
Stopping rule	no improvement in post-verification consensus confidence

A.3 Inference information boundary

Reference answers and post-hoc claim labels are never exposed to the test-time coordinator. Proposers can use only the task input, their assigned role, and role-permitted observations. Candidate extraction carries forward structured claim spans and source evidence from proposer outputs. Verifiers receive only selected claim packets and the evidence needed to check those claims. Both routing controls reuse DASH-Agent's initial responses and hard verification-slot budget, but not the same candidate pool. Random routing samples uniformly from DASH-Agent's graph-exposed $\mathcal{C}^t$, isolating triage, whereas reliability-only routing bypasses graph discovery and ranks the structured claims whose answer implication explicitly disagrees with the current majority using $R(c)$ alone. This distinction is intentional: the latter control tests reliability without graph-conditioned exposure.

A.4 Operational graph, claim, and reliability construction

This subsection makes explicit the fixed operations between the structured proposer output and DST. No test labels, learned edge predictor, or auxiliary free-form claim generator is used on this path.

Interaction graph. Let h_q, h_i^m, and h_i^e denote unit-normalized frozen-backbone embeddings of the query, agent-i message, and linked evidence text. Cosine similarities are mapped from $[-1, 1]$ to $[0, 1]$. For pair (i, j), the four symmetric cues are

$$\begin{aligned} r_{ij}^t &= \tfrac{1}{2}\left[s(h_i^m, h_q) + s(h_j^m, h_q)\right], & p_{ij}^t &= \mathbf{1}[r_i \neq r_j], \\ c_{ij}^t &= \mathbf{1}[a_i^t \neq a_j^t], & e_{ij}^t &= s(h_i^e, h_j^e), \end{aligned} \tag{5}$$

where $s(u, v) = (1 + u^\top v)/2$. If either response lacks linked evidence, e_{ij}^t takes the neutral value 0.5. The pre-sparsification edge weight is

$$\bar{A}_{ij}^t = \tfrac{1}{4}\left(r_{ij}^t + p_{ij}^t + c_{ij}^t + e_{ij}^t\right), \qquad \bar{A}_{ii}^t = 0. \tag{6}$$

The final A^t keeps the union of the top-K_s neighbors of every agent and uses the corresponding symmetric weight. The graph therefore contains no separately trained edge scorer.

Claim mapping and admission. The proposer specification in Appendix F requires a structured list $\mathcal{Q}_i^t = \{c_{i\ell}\}$ of claim records with evidence-source identifiers and answer implications. Candidate extraction operates on these records rather than asking another language model to rediscover spans.

For source agent i, graph novelty is summarized by

$$\delta_i^t = \frac{\|x_{\mathrm{dis},i}^t\|_2}{\|x_{\mathrm{dis},i}^t\|_2 + \|x_{\mathrm{con},i}^t\|_2 + 10^{-8}}. \tag{7}$$

The source score is then refined at claim level. Let $d_{\mathrm{ans}}(c_{i\ell}) \in \{0, 1\}$ indicate whether the claim's answer implication conflicts with the current neighborhood majority. Let

$$d_{\mathrm{ev}}(c_{i\ell}) = 1 - \max_{c' \in \mathcal{Q}_{\mathcal{N}(i)}^t} s(h_{c_{i\ell}}, h_{c'}), \qquad s_{\mathrm{disc}}(c_{i\ell}) = \delta_i^t \, \frac{d_{\mathrm{ans}}(c_{i\ell}) + d_{\mathrm{ev}}(c_{i\ell})}{2}, \tag{8}$$

where h_c is the frozen embedding of the claim together with its linked evidence and $\mathcal{Q}_{\mathcal{N}(i)}^t$ contains claims emitted by graph neighbors. If a claim has no explicit answer implication, $d_{\mathrm{ans}} = 0$ and evidence divergence can still expose it. The candidate set is

$$\mathcal{C}^t = \{c_{i\ell} \in \cup_i \mathcal{Q}_i^t : s_{\mathrm{disc}}(c_{i\ell}) \geq \tau_d\}, \tag{9}$$

with τ_d selected on held-out validation data and frozen at test time. Claims from one agent share δ_i^t but can differ in d_{ans} and d_{ev}, so graph variation conditions exposure without admitting every claim from a high-variation speaker. Claim-specific utility and reliability are computed only after admission and do not determine membership in $\mathcal{C}^t$.

Reliability score and held-out calibration. For candidate claim c, $q_{\mathrm{con}}(c)$ is the claim–evidence cosine similarity mapped to $[0, 1]$, $q_{\mathrm{tool}}(c)$ is 1 for a schema-valid executed tool result and 0 for a failed or invalid result, $q_{\mathrm{conf}}(c)$ is the reported confidence, and $q_{\mathrm{prior}}(c)$ is the source agent's held-out correctness rate. The tool cue is omitted on non-tool tasks rather than set to zero, giving

$$R_{\mathrm{raw}}(c) = \frac{1}{|\mathcal{K}_c|} \sum_{k \in \mathcal{K}_c} q_k(c), \qquad R(c) = g_{\mathrm{cal}}(R_{\mathrm{raw}}(c)), \tag{10}$$

where $\mathcal{K}_c$ is the set of available cues and g_{cal} is an isotonic-regression map fitted once on held-out evidence labels and then frozen. Calibration labels are used only to fit g_{cal} and are never supplied to discovery, DST, or verification on the test split. This separation prevents the reliability term from becoming an unreported correctness oracle.

A.5 Sensitivity of the selected coordinator configuration

Figure A.1 tests whether the validation-selected operating point depends on a narrow hyperparameter choice. We vary four coordinator parameters around Table A.1: graph neighborhood size K_s, candidate-routing cap ρ, maximum rounds T, and DST reliability weight w_R. For each seed, benchmark scores are normalized by the five-seed mean of the selected DASH-Agent setting, then averaged across the five primary benchmarks and multiplied by 100. The selected setting therefore has mean 100 while preserving seed-level variation. When w_R changes, $w_N : w_U : w_C$ remains $0.30 : 0.25 : 0.10$ and the three weights are renormalized to sum to $1 - w_R$.

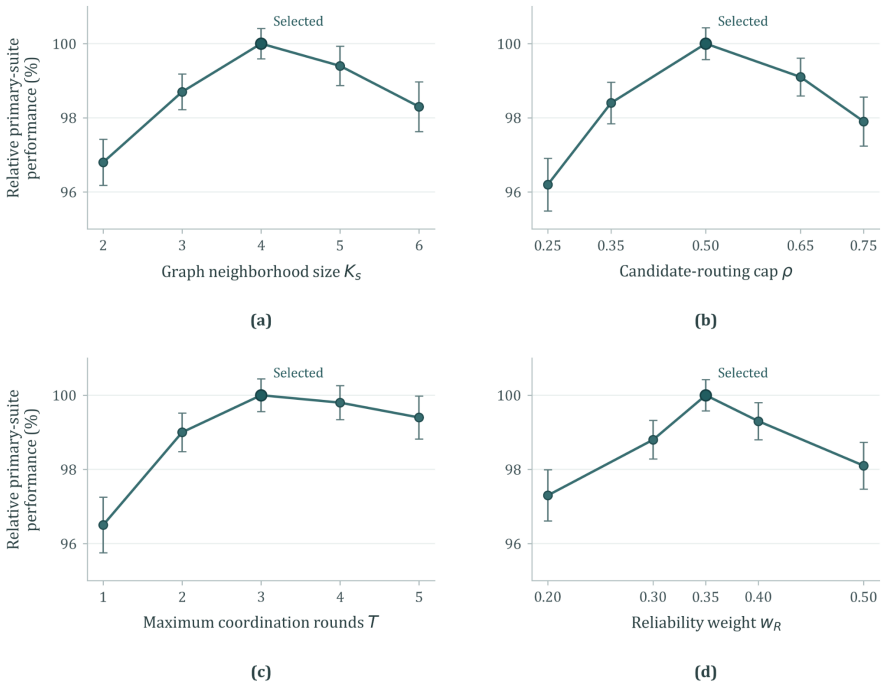

Figure A.1: Coordinator sensitivity around the validation-selected setting. Panels vary (a) K_s, (b) routing cap ρ, (c) maximum rounds T, and (d) reliability weight w_R. Values are normalized to mean 100 at the selected setting, and error bars denote standard deviation over five seeds.

Across all four sweeps, performance changes smoothly around the selected point rather than peaking at an isolated value. Smaller neighborhoods and tighter routing caps reduce useful candidate coverage, whereas larger values admit more weakly related or lower-priority exchanges. Additional rounds plateau after three. Likewise, too little reliability weight weakens evidence-aware prioritization, while excessive weight can down-rank useful minority claims with only moderate prior support. The broad local plateaus support the fixed configuration without suggesting that the reported gains depend on a narrow optimum.

B Theoretical Analysis

The main paper motivates DASH-Agent as a decision decomposition rather than a stack of interchangeable heuristics. This section formalizes three properties behind that interpretation. The analysis does not equate graph variation with correctness. It establishes what the graph filter measures, why downstream routing cannot repair missed exposure, and why claim-level decisions are strictly more expressive than speaker-level gating when a response mixes claims of different quality.

B.1 Graph variation characterizes local disagreement

Assume the round-t interaction graph is symmetric and every retained node has positive degree. Let A^t be its adjacency matrix, D^t its degree matrix, and

$$L^t = I - (D^t)^{-1/2} A^t (D^t)^{-1/2} \tag{11}$$

be the symmetric normalized graph Laplacian. For readability, the superscript t is omitted in the remainder of this subsection.

Proposition 1. Let $L = \Phi \Lambda \Phi^\top$ with eigenvalues $\lambda_j \in [0, 2]$, and define $X_{\mathrm{dis}} = \frac{1}{2} L X$. Then

$$\|X_{\mathrm{dis}}\|_F^2 = \frac{1}{4} \sum_j \lambda_j^2 \|\phi_j^\top X\|_2^2, \tag{12}$$

and the normalized Dirichlet energy is

$$\operatorname{tr}(X^\top L X) = \frac{1}{2}\sum_{i,j} A_{ij} \left\| \frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right\|_2^2. \tag{13}$$

Consequently, X_{dis} places larger spectral weight on components with larger λ_j, which represent stronger local variation under the current interaction graph. Moreover,

$$0 \le \|X_{\mathrm{dis}}\|_F^2 \le \frac{1}{2}\operatorname{tr}(X^\top L X). \tag{14}$$

Proof. Substituting $L = \Phi\Lambda\Phi^\top$ into $X_{\mathrm{dis}} = \frac{1}{2}LX$ gives

$$X_{\mathrm{dis}} = \frac{1}{2}\Phi\Lambda\Phi^\top X. \tag{15}$$

Because Φ is orthonormal, the squared Frobenius norm is invariant to left multiplication by Φ, which gives Equation 12. Equation 13 follows from the quadratic form of the symmetric normalized Laplacian applied column-wise to X. Finally, $\lambda_j \in [0, 2]$ implies $\lambda_j^2/4 \le \lambda_j/2$. Applying this inequality to each spectral component yields Equation 14. □

Proposition 1 specifies the role of the graph filter without assigning semantic correctness to high-frequency content. Larger disagreement-component energy means that the corresponding agent state contains stronger variation relative to its graph neighborhood. Whether an associated claim is useful or correct remains a separate validation and verification question.

B.2 Discovery and downstream recovery are non-substitutable

Let $\mathcal{Q}^t$ be the claims available in agent messages at round t, $\mathcal{C}^t \subseteq \mathcal{Q}^t$ the claims exposed by discovery, $\mathcal{V}^t \subseteq \mathcal{C}^t$ the claims selected for verification, and $\mathcal{S}^t \subseteq \mathcal{V}^t$ the claims returned as `supported`. Let $\mathcal{U}^t \subseteq \mathcal{Q}^t$ denote the useful dissent claims under post-hoc evaluation.

Proposition 2. For every round t,

$$\frac{|\mathcal{U}^t \cap \mathcal{S}^t|}{|\mathcal{U}^t|} \le \frac{|\mathcal{U}^t \cap \mathcal{C}^t|}{|\mathcal{U}^t|}, \tag{16}$$

whenever $|\mathcal{U}^t| > 0$. Thus, downstream prioritization or verification cannot recover a useful claim that discovery does not expose.

Proof. The pipeline construction gives $\mathcal{S}^t \subseteq \mathcal{V}^t \subseteq \mathcal{C}^t$. Intersecting all sets with $\mathcal{U}^t$ preserves inclusion, so $\mathcal{U}^t \cap \mathcal{S}^t \subseteq \mathcal{U}^t \cap \mathcal{C}^t$. Taking cardinalities and dividing by the positive denominator $|\mathcal{U}^t|$ gives Equation 16. □

The right-hand side is candidate recall over useful claims, which is therefore an upper bound on useful-claim recovery through the downstream verification path. This is why discovery and reliability-only routing diagnose different failure modes. A stronger reliability estimator can improve selection among exposed claims, but it cannot recover a claim that never enters $\mathcal{C}^t$.

B.3 Claim-level coordination is strictly more expressive

For agent i, let $\mathcal{Q}_i$ be the set of claims in its response. A binary speaker-level policy assigns one decision $g_i \in \{0, 1\}$ to every claim from agent i. A claim-level policy assigns a decision $g_{ic} \in \{0, 1\}$ separately for each $c \in \mathcal{Q}_i$.

Proposition 3. The class of binary speaker-level keep/drop policies is a strict subset of the class of claim-level keep/drop policies whenever at least one response can contain two claims that require different decisions.

Proof. Every speaker-level policy is realizable by a claim-level policy by setting $g_{ic} = g_i$ for all $c \in \mathcal{Q}_i$, so claim-level policies are at least as expressive. Now consider a response containing a supported claim c^+ that should be retained and an unsupported claim c^- that should be rejected. The desired assignment is $g_{ic^+} = 1$ and $g_{ic^-} = 0$. A binary speaker-level gate has only two choices. Setting $g_i = 1$ retains both claims, while setting $g_i = 0$ rejects both. It cannot realize the mixed assignment, whereas a claim-level policy can. The containment is therefore strict. □

This result formalizes why DASH-Agent keeps the downstream unit at claim level even though discovery begins from agent states. The benefit is not that every claim must be verified independently. It is that the coordinator can express different outcomes when one speaker mixes useful evidence with unsupported inference.

B.4 Priority normalization and budgeted selection

For DST score S_i^t, the normalized priority is

$$\pi_i^t = \frac{\exp(S_i^t)}{\sum_{j \in \mathcal{C}^t} \exp(S_j^t)}. \tag{17}$$

For any two candidates i and j, $\pi_i^t > \pi_j^t$ if and only if $S_i^t > S_j^t$ because the exponential function is strictly increasing and the denominator is common within the round. Softmax therefore normalizes relative priority without changing its ordering.

For a generic heterogeneous verification cost b_i^t, the routing rule is

$$\mathcal{V}^t = \arg\max_{\mathcal{S} \subseteq \mathcal{C}^t} \sum_{i \in \mathcal{S}} \pi_i^t \quad \text{s.t.} \quad \sum_{i \in \mathcal{S}} b_i^t \le B_t, \qquad |\mathcal{S}| \le \left\lfloor \rho |\mathcal{C}^t| \right\rfloor, \tag{18}$$

where $\rho = 0.5$ in the reported configuration. Priority ordering alone does not imply a greedy solution when the b_i^t values differ. In the reported implementation, however, the hard budget is measured in verification slots and each length-capped claim packet has unit slot cost $b_i^t = 1$. Equation 18 then reduces exactly to selecting the $K_t = \min\{B_t, \lfloor \rho |\mathcal{C}^t| \rfloor\}$ highest-priority claims, which is the rule used in Algorithm 1.

B.5 Computational complexity

For N agents with d-dimensional states, pairwise graph compatibility costs $O(N^2 d)$. Sparse graph filtering costs $O(|E|d)$ for retained edge set E. Let $M = |\mathcal{C}^t|$ be the number of candidate claims in a round. Computing agent–claim compatibility before thresholding costs $O(NMd)$. With unit verification-slot costs, selecting the top K_t priorities can be implemented in $O(M \log M)$ time by sorting, or $O(M)$ expected time with a selection routine. If verification group H_k exchanges pairwise evidence for selected issue k, the interaction term is bounded by $O(|H_k|^2)$. The non-generation coordination cost per round is therefore

$$O\left(N^2 d + |E| d + NMd + M \log M + \sum_{k \in \mathcal{V}^t} |H_k|^2 \right). \tag{19}$$

Language-model decoding and external tool latency are not represented by this symbolic bound and are measured separately through exchanged-token accounting.

C Evaluation Protocol and Metric Definitions

C.1 Benchmarks, seeds, and fairness controls

The primary suite contains GSM8K and MATH for mathematical reasoning, HotpotQA for multi-hop question answering, and API-Bank and ToolBench for tool use. The task metrics are accuracy for GSM8K and MATH, answer F1 for HotpotQA, and task success for API-Bank and ToolBench. Every reported primary mean and standard deviation is computed over five random seeds. Table 1 fixes Qwen2.5-7B-Instruct across methods so that the primary comparison changes the coordination policy rather than model capacity.

Random routing and reliability-only routing reuse the initial responses and hard verification-slot budget used by DASH-Agent. Random routing draws uniformly from the graph-exposed candidate set, while reliability-only routing bypasses graph discovery and ranks explicit answer-dissent claim records using calibrated reliability alone. The validation-selected coordinator is frozen at test time. Reference answers and post-hoc evidence labels are used only for evaluation and never for routing.

C.2 Claim audit and preservation metrics

The mechanism audit is fixed before inspecting any routing outcome. We sample 240 primary-test instances per benchmark from instances that contain at least one structured dissent claim, giving 1,200 audited instances in total. Sampling is uniform within benchmark and seed, and the same audited instances are reused for all routing comparisons. The initial majority is fixed from the initial answers. Each dissent record contains the structured claim text, source-agent identifier, evidence pointer, answer implication, and a post-hoc evidence disposition. Support is checked against task-specific independent evidence: an independently checked derivation for mathematical reasoning, supporting passages for HotpotQA, and executed return values plus schema constraints for API-Bank and ToolBench. Routing decisions, final answers, and method identity are hidden during annotation.

Two annotators independently label every audited claim. On the three-way evidence disposition (*supported*, *contradicted*, *unresolved*), raw agreement is 88.6% with Cohen's $\kappa = 0.81$. On whether a supported claim can materially affect the reference resolution, agreement is 85.9% with $\kappa = 0.77$. A third annotator adjudicates disagreements using the same evidence packet and remains blind to routing outcome. Unresolved claims are not forced into the useful- or incorrect-dissent sets. A useful dissent claim must be supported and materially relevant to the reference resolution, while an incorrect dissent claim must be contradicted by the independent evidence. A claim counts as preserved only if it remains available to final synthesis. Mere appearance in the transcript is insufficient.

Table C.1 summarizes the audit scale. The 1,200 instances yield 3,828 dissent claims, including 1,284 useful claims and 1,796 incorrect claims. The remaining 748 claims are unresolved or supported but non-decisive and are excluded from UDS/IDS denominators. Among the audited instances, 512 satisfy the MCR eligibility condition of an incorrect initial majority with at least one evidence-supported corrective minority claim.

Table C.1: Scale of the claim-level mechanism audit. Counts are fixed before routing comparisons.

Benchmark	Instances	Dissent claims	Useful	Incorrect
GSM8K	240	682	222	326
MATH	240	755	248	361
HotpotQA	240	881	314	402
API-Bank	240	734	231	349
ToolBench	240	776	269	358
Total	**1,200**	**3,828**	**1,284**	**1,796**

Let $\mathcal{I}_{\text{elig}}$ be the set of instances with an incorrect initial majority and at least one evidence-supported correct minority claim. Let $\mathcal{U}$ be the set of annotated useful dissent claims and $\mathcal{W}$ the set of annotated incorrect dissent claims. Let $\text{keep}(c) = 1$ when claim c remains available to final synthesis and $\text{drop}(c) = 1 - \text{keep}(c)$. The three mechanism metrics are

$$\text{MCR} = \frac{1}{|\mathcal{I}_{\text{elig}}|} \sum_{x \in \mathcal{I}_{\text{elig}}} \mathbf{1}\left[\exists c \in \mathcal{U}_x : \text{keep}(c) = 1\right], \tag{20}$$

$$\text{UDS} = \frac{\sum_{c \in \mathcal{U}} \text{keep}(c)}{|\mathcal{U}|}, \qquad \text{IDS} = \frac{\sum_{c \in \mathcal{W}} \text{drop}(c)}{|\mathcal{W}|}. \tag{21}$$

MCR is instance-level, while UDS and IDS are claim-level and use different populations. Their unweighted average is used only as a compact visualization aid in Figure 4(b) and is not treated as an independent evaluation metric.

C.3 Reliability calibration

Let $R_i \in [0,1]$ be the pre-verification reliability estimate for candidate claim i and $y_i \in \{0,1\}$ its post-hoc evidence correctness label. For ten score intervals $\{B_m\}_{m=1}^{10}$, Expected Calibration Error (ECE) is

$$\mathrm{ECE} = \sum_{m=1}^{10} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \tag{22}$$

where $\mathrm{acc}(B_m)$ is the empirical correctness fraction and $\mathrm{conf}(B_m)$ is the mean reliability estimate inside the interval. The Brier score is

$$\mathrm{Brier} = \frac{1}{n}\sum_{i=1}^{n}(R_i - y_i)^2. \tag{23}$$

The main-paper calibration values are ECE 0.053 ± 0.007 and Brier score 0.087 ± 0.009. Both are computed seed-wise from individual claim scores after the held-out calibration map is frozen. Table C.2 additionally reports pooled bin support and the pooled mean score. The support counts aggregate candidate claims across the five test seeds, while correctness retains the seed-wise mean and standard deviation used in the main paper. Because candidate generation can differ by seed, the pooled counts are descriptive rather than five copies of one fixed claim set.

Using the pooled supports and rounded mean scores in Table C.2 gives ECE 0.053 after rounding, consistent with the reported seed-wise mean. The Brier score is computed from individual scores and binary labels and therefore cannot be reconstructed from bin means alone. The low-score intervals are visibly conservative rather than uniformly calibrated. Their contribution to aggregate ECE is limited because 70% of evaluated candidates fall in the two highest score intervals. We therefore use ECE as a distribution-weighted summary and rely on the full bin table to expose region-specific calibration behavior.

Table C.2: Calibration-bin support, mean reliability score, and observed correctness. Support is pooled across five test seeds, while correctness is percent, mean $\pm$ standard deviation over seeds.

Score interval	Support	Mean score (%)	Correctness (%)
$[0.0, 0.1)$	96	6.1	29.8 ± 3.4
$[0.1, 0.2)$	176	16.9	37.6 ± 2.7
$[0.2, 0.3)$	258	25.7	46.9 ± 3.0
$[0.3, 0.4)$	386	36.6	55.3 ± 2.1
$[0.4, 0.5)$	548	46.3	63.8 ± 2.5
$[0.5, 0.6)$	932	56.9	71.7 ± 1.8
$[0.6, 0.7)$	1,453	66.1	78.9 ± 1.6
$[0.7, 0.8)$	2,204	76.8	85.2 ± 1.7
$[0.8, 0.9)$	4,942	86.6	90.8 ± 1.1
$[0.9, 1.0]$	8,851	95.2	94.6 ± 0.9
Total	**19,846**	–	–

D Extended Numerical Results

This section reports the numerical values underlying the main-paper mechanism figures and derived ablations. These tables support reproducibility and do not define additional evaluation targets.

D.1 Claim-level preservation and discovery

Table D.1 reproduces the values used in Figure 3(a), while Table D.2 gives the discovery and triage values in Figure 3(b). Together they preserve the distinction emphasized in the main text: graph filtering broadens candidate exposure, while adding DST primarily improves useful precision.

Table D.1: Claim-level preservation and suppression (percent, mean ± standard deviation over five seeds).

Method	MCR↑	UDS↑	IDS↑
Majority Voting	43.6 ± 2.4	39.8 ± 2.1	53.1 ± 1.8
Debate	65.2 ± 1.7	61.4 ± 1.9	45.0 ± 2.2
Graph-of-Agents	70.4 ± 1.5	66.8 ± 1.6	58.9 ± 1.7
AFlow	69.1 ± 1.8	65.7 ± 1.4	59.6 ± 1.9
AgentPrune	63.0 ± 2.0	60.3 ± 1.7	64.2 ± 1.6
Random Routing	56.8 ± 2.3	53.1 ± 2.5	50.4 ± 2.1
Reliability-only Routing	$\underline{75.2 \pm 1.2}$	$\underline{72.4 \pm 1.4}$	$\underline{69.7 \pm 1.3}$
DASH-Agent	$\mathbf{84.7 \pm 0.9}$	$\mathbf{81.3 \pm 1.1}$	$\mathbf{78.2 \pm 1.0}$

Table D.2: Candidate discovery and triage (percent, mean ± standard deviation over five seeds).

Method	Candidate recall↑	Useful precision↑
Random disagreement	43.8 ± 2.6	37.5 ± 2.8
Confidence gap	56.9 ± 2.1	50.7 ± 2.3
Semantic distance	62.7 ± 1.9	58.4 ± 1.8
Graph spectral filtering	$\mathbf{74.8 \pm 1.5}$	$\underline{64.1 \pm 1.7}$
Filtering + DST	$\underline{73.6 \pm 1.3}$	$\mathbf{82.7 \pm 1.1}$

D.2 Task-level ablation values

Table D.3 reconstructs the task means behind Figure 4(a) from the full DASH-Agent scores and the plotted percentage-point drops. Because Figure 4(a) reports differences rather than seed-level paired observations, the table reports means only and does not introduce uncertainty for the derived variants.

Table D.3: Task means reconstructed from the full DASH-Agent scores and the percentage-point drops shown in Figure 4(a). Average is the unweighted mean of GSM8K, MATH, and HotpotQA.

Variant	GSM8K	MATH	HotpotQA	Average
Full DASH-Agent	**83.4**	**47.6**	**72.9**	**68.0**
w/o candidate dissent discovery	81.4	46.2	70.6	66.1
w/o DST reliability validation	79.6	45.4	68.8	64.6
w/o selective verification	80.8	45.7	70.0	65.5
Reliability-only routing	78.7	44.3	67.7	63.6
Random routing	77.1	42.8	66.0	62.0

The ablations should be interpreted sequentially. Discovery controls the feasible claim set, DST allocates scarce checking capacity within that set, and verification decides which selected claims survive as supported evidence. Their drops are interventions on one path and should not be added as independent component contributions.

D.3 Communication accounting

Let T_M denote all exchanged tokens for method M over the same benchmark suite. Accounting includes initial generation, routed evidence, verification, judging, synthesis, and tool exchanges where applicable. The horizontal axis in Figure 5 is a shared cost index. Dividing each plotted cost by Debate's value 2.5 gives the method's communication ratio relative to Debate. Table D.4 reports both quantities together with S_{norm}.

Table D.4: Communication–performance values underlying Figure 5 (lower cost, higher S_{norm}).

Method	Cost index↓	Cost / Debate↓	$\mathbf{S}_{\mathrm{norm}}$ ↑
Debate	2.5	1.0	92.6
Graph-of-Agents	1.8	0.7	94.5
AFlow	1.9	0.8	94.0
AgentPrune	**1.3**	**0.5**	93.2
ReConcile	1.6	0.6	95.0
CONSENSAGENT	1.8	0.7	95.7
SELENE	2.0	0.8	<u>96.3</u>
DASH-Agent	<u>1.5</u>	<u>0.6</u>	**100.0**

The normalized performance score first divides each method's score on each benchmark by the best score on that benchmark, then averages the five ratios. It is therefore a cross-task display score rather than the arithmetic mean of the five raw benchmark values. DASH-Agent uses $1.5/2.5 \approx 0.6$ times Debate's exchanged-token cost at the operating point shown in Figure 5.

D.4 Backbone and additional-benchmark transfer

The primary comparison fixes Qwen2.5-7B-Instruct. The transfer checks stated in the main paper use Llama-3.1-8B-Instruct and include GSM8K, CommonsenseQA, and StrategyQA. Table D.5 reports only values and margins stated in the main text. Matched uncertainty is not available for the two auxiliary benchmarks and is therefore not inferred here.

Table D.5: Transfer results stated in the main paper. The gap column is reported only where a matched strongest-alternative margin is available.

Evaluation	Backbone	DASH-Agent	Reported gap
GSM8K	Llama-3.1-8B-Instruct	80.2 ± 0.7	–
CommonsenseQA	Llama-3.1-8B-Instruct	78.6	+1.3
StrategyQA	Llama-3.1-8B-Instruct	74.4	+1.5

E Illustrative Case Traces

Table E.1: Two illustrative claim-level traces. The first preserves corrective minority evidence, while the second prevents unsupported dissent from becoming a final fact.

Stage	Corrective minority	Unsupported dissent
Initial state	Majority dismisses an upgrade because a risky feature is disabled. A minority claim notes that the installed dependency version remains affected.	A minority claim says an API operation succeeded because the proposed call is syntactically valid.
Discovery	The version claim conflicts with the local majority and links to a concrete package manifest.	The execution claim conflicts materially with the majority state.
DST	Direct manifest support and decision relevance yield high utility and reliability.	Syntax alone does not evidence execution, so reliability remains low.
Verification	The affected version is confirmed, while the disabled feature lowers immediate exploitability.	No status, returned object, or execution record supports success, so the verifier returns `unresolved`.
Synthesis	Recommend the upgrade without claiming current exploitability.	Do not report completion. The dissent remains unresolved rather than becoming a fact.

F PROMPT TEMPLATES

The following implementation specifications define the model-facing inputs, required fields, and decision constraints. Bracketed fields are filled from the current task or selected verification packet.

F.1 PROPOSER SPECIFICATION

Purpose	Produce an independent solution with inspectable claims and evidence.
Inputs	`[QUERY]`, `[ROLE]`, `[ROLE-SPECIFIC OBSERVATIONS]`.
Task	Give a provisional answer and the minimal supporting claims. For each claim, identify its evidence source and distinguish observation from inference.
Required output	`answer`, `claims`, `evidence sources`, `confidence`, `unresolved assumptions`.
Constraints	Do not infer agreement from peers or report a tool result without an execution record. Confidence must lie in $[0, 1]$.

F.2 VERIFIER SPECIFICATION

Purpose	Adjudicate one selected claim from evidence rather than novelty or agreement.
Inputs	`[QUERY]`, `[CANDIDATE CLAIM]`, `[SOURCE EVIDENCE]`, `[RELEVANT COUNTEREVIDENCE]`.
Task	Check the decisive derivation, passage, schema condition, or executed tool observation independently.
Required output	`disposition` $\in$ {`supported`, `contradicted`, `unresolved`}, `decisive evidence`, `remaining uncertainty`.
Constraints	Novelty and majority agreement are not evidence of correctness. Return `unresolved` when the available evidence cannot decide the claim.

F.3 SYNTHESIS SPECIFICATION

Purpose	Produce the final answer from initial responses and verified claim records.
Inputs	`[QUERY]`, `[INITIAL RESPONSES]`, `[VERIFIED CLAIM RECORDS]`.
Task	Use supported evidence, retain a verified minority claim when decisive, exclude contradicted claims, and keep unresolved claims explicitly uncertain.
Required output	`final answer`, `decisive evidence`, `remaining uncertainty`, `consensus confidence`.
Constraints	Do not restore a contradicted claim because it is popular or convert an unresolved claim into fact.