COUNTERFACTUAL WITNESS EXCHANGE FOR MULTI-AGENT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent question answering can produce repeated agreement from the same evidence. In multi-hop questions, examining passages individually does not resolve this problem because several passages may be needed together. We introduce Counterfactual Witness Exchange (CWE), which evaluates an answer by the largest number of supporting evidence sets that do not share a passage. It tests each set by restarting the answering process with only that evidence. To allocate a limited inference budget, we derive a query rule that maximizes the guaranteed one-step reduction in capacity uncertainty. Recomputing the supporting sets after each replay allows an earlier witness to be replaced by several compatible alternatives. Experiments on HotpotQA and 2WikiMultiHopQA show improved answer quality and lower token use compared with repeated full-context inference. Controlled comparisons examine how accounting for overlap and discovering alternative supporting sets contribute to answer selection.

1 INTRODUCTION

Multi-agent debate lets models compare responses and revise their answers through discussion (Du et al., 2024; Chen et al., 2024; Smit et al., 2024). For questions answered from retrieved passages, several agents may nevertheless agree because they rely on the same evidence. Additional agreement can then repeat the support already available. Our accompanying communication study finds that discussion can strengthen agreement on wrong answers (Appendix C). Selecting an answer therefore requires examining the evidence behind the agreement.

Multi-hop questions make this difficult. Two passages may support an answer together even when neither suffices alone. Testing individual passages can miss this complementary information. Counting every supporting combination creates a different problem: many combinations may all require the same passage. A fixed partition avoids such reuse but may separate passages needed together. The goal is to find complete supporting sets and determine how many can be retained without counting a passage more than once.

We introduce *Counterfactual Witness Exchange* (CWE) to pursue this goal under a limited inference budget. CWE tests a set by restarting the entire answering process using only the selected evidence. We call a set that reproduces an answer a *witness*. The maximum number of witnesses that share no passage defines the answer's *support capacity*. This objective keeps complementary passages together while accounting for reuse across witnesses. It makes the overlap between supporting sets part of the answer-selection decision.

Finding these witnesses creates an acquisition problem: which evidence subset should receive the next replay? CWE maintains an interval for each candidate's capacity and chooses the query that leaves the smallest total interval width under its worst possible response. We derive the optimal one-step rule by computing how support for, or against, each candidate changes its capacity bounds. A query is valuable when it can establish additional support or eliminate a possible witness.

These effects depend on the full overlap structure. An early witness can conflict with two later witnesses that are disjoint from each other. Replacing it with both gives a larger compatible collection. CWE recomputes the packings after every replay, allowing these exchanges to change both answer scores and subsequent query choices. Figure 1 illustrates the procedure.

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.

We evaluate CWE on HotpotQA and 2WikiMultiHopQA with multiple model backbones and decoding seeds. The comparison includes repeated full-context inference and alternative policies for querying evidence subsets. CWE improves answer quality over repeated full-context inference while using fewer tokens. The mechanism experiments separate scoring from search. With replay observations held fixed, capacity scoring yields higher accuracy than counting supporting sets. A matched query restriction further shows that searching for alternative combinations is more valuable when an early witness obstructs them. These comparisons connect the method's behavior to the support structure it is designed to recover.

Our contributions are as follows:

- **A support objective that accounts for evidence reuse.** We formulate support capacity as the largest collection of complete supporting sets that share no evidence. The objective preserves complementary information within each set and accounts for overlap across sets.
- **An optimal one-step acquisition rule.** We formulate query selection as minimizing worst-case capacity uncertainty and derive an explicit rule from the sensitivity of packing bounds. Recomputing these bounds allows newly discovered witnesses to replace earlier selections.
- **An empirical separation of scoring and search.** We compare answer rules on identical replay observations and use matched support structures to test when querying alternative combinations improves selection. These analyses connect the overall results to evidence overlap and witness exchange.

2 Related Work

Deliberation and dependent support. Multi-agent debate revises answers through peer discussion (Du et al., 2024), while ReConcile combines heterogeneous models through confidence-weighted voting (Chen et al., 2024). Comparisons with cost-matched ensembles separate communication from additional inference (Smit et al., 2024). Analyses of debate dynamics attribute its gains to initial opinion diversity and communicated confidence (Zhu et al., 2026b), and failure taxonomies trace many multi-agent breakdowns to inter-agent misalignment and missing verification (Cemri et al., 2025). Free-MAD scores answer trajectories without requiring final consensus (Cui et al., 2026), and MAD-M^2 masks shared memories to limit error propagation (Tian et al., 2026). Debate protocols have also been applied to misinformation detection (Han et al., 2025). Correlated model errors limit the information supplied by agreement (Kim et al., 2025). Models can conform to an incorrect majority (Zhu et al., 2025), and user-facing assistants can align with stated beliefs rather than truth (Sharma et al., 2024); a social-psychology account links such outcomes to how agent interactions are organized (Zhang et al., 2024). Repeated information has long been modeled as a source of correlated judgment: informational cascades can outweigh private signals (Bikhchandani et al., 1992), and persuasion bias describes the failure to discount repetition (DeMarzo et al., 2003). Dependence-aware truth discovery adjusts votes for inferred source copying (Dong et al., 2009); CONFER discounts confidence when logical derivations share premises (Tammet et al., 2021). CWE measures support through evidence-subset replays. It uses overlap between supporting sets to determine which can count together, linking reused evidence to answer scoring and query selection.

Evidence isolation and attribution. RobustRAG partitions passages into disjoint groups, generates isolated responses, and aggregates keywords or next-token distributions (Xiang et al., 2026). Its certificates concern retrieval corruption. CWE instead queries overlapping subsets and imposes disjointness when counting support. This allows the selected grouping to change when fixed groups separate complementary passages or an early witness obstructs alternatives. ERASER evaluates rationale sufficiency and comprehensiveness by retaining or removing evidence (DeYoung et al., 2020). ContextCite fits a sparse surrogate to response probabilities under context ablations (Cohen-Wang et al., 2024), while Context-Picker learns compact evidence selection from reference-answer supervision and repeated leave-one-out pruning (Zhu et al., 2026a). Answer quality can also depend on where relevant information appears within a supplied context (Liu et al., 2024), and ClashEval perturbs retrieved content to measure the tug-of-war between a model's prior and external evidence (Wu et al., 2024). GAVEL debates atomic claims under explicit evidence contracts with mechanized citation checks (Xu et al., 2026); atomic-fact evaluation decomposes long-form answers into individually checkable claims (Min et al., 2023), and lightweight NLI checkers trace nonfactual output back to

context chunks (Sankararaman et al., 2024). CWE maintains several supporting sets per answer and scores their compatibility. It can retain a set whose passages work only jointly, while preventing any one passage from contributing to multiple counted supports for the same answer.

Budgeted search over evidence. Chain-of-thought prompting elicits intermediate reasoning steps (Wei et al., 2022). Self-consistency aggregates sampled reasoning paths (Wang et al., 2023), Tree of Thoughts searches partial solutions (Yao et al., 2023), and test-time scaling studies how difficulty and budget govern inference allocation (Snell et al., 2025). SELENE selectively triggers debate and judges existing responses under paraphrased or subset-sampled evidence (Verma et al., 2026). CWE regenerates the complete answering trajectory for each evidence subset. Each replay can reveal a new answer or a new supporting set. Its query rule minimizes worst-case capacity uncertainty over one replay, and recomputing the witness collection allows compatible alternatives to replace an earlier support. This connects acquisition to the support objective used for final answer selection.

3 METHOD

CWE connects answer selection and evidence acquisition through two objectives. The first measures support by the largest collection of complete witnesses that share no evidence. The second chooses the next replay to minimize the worst-case uncertainty remaining about these capacities. We derive an exact one-step query rule from the effect of each possible response on the capacity bounds. Recomputing the packings after every replay permits witness exchange. Figure 1 shows the procedure.

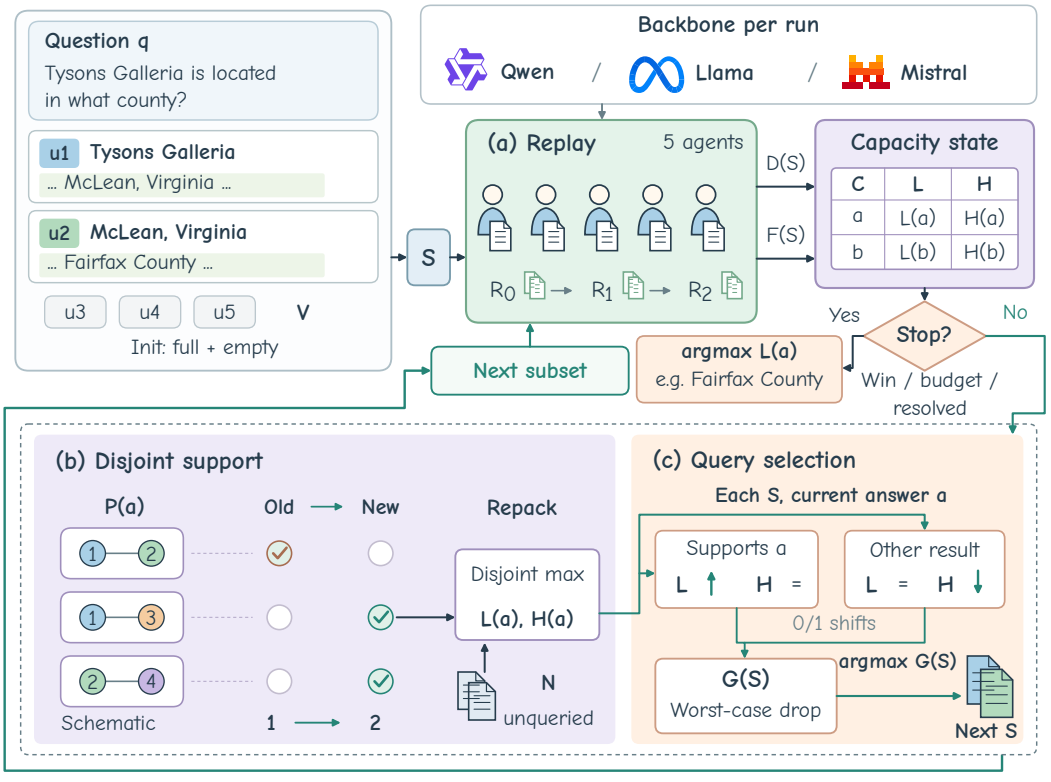

Figure 1: **Counterfactual Witness Exchange.** (a) Clean replays discover candidate answers and supporting sets. (b) Disjoint packing measures capacity; exchange can replace one witness with two compatible witnesses for the same answer. (c) Positive and negative response sensitivities determine each query's worst-case reduction in capacity uncertainty. CWE maximizes this guaranteed reduction and returns the selected subset to replay. Passage excerpts come from a HotpotQA question; replay outputs and witness exchange are illustrative.

3.1 PRELIMINARIES AND NOTATION

For question q, let $V = \{u_1, \ldots, u_m\}$ be the distinct visible evidence texts, including titles. Exact copies map to one unit. Units receive fixed canonical IDs and a fixed text order; each appears once in a subset prompt.

To test $S \subseteq V$, we restart five agents at R0 using only S. At R1 and R2, they receive the preceding round's parsed answers from this same replay. Restarting prevents a removed passage from returning through an earlier peer report. The model, prompt, parser, decoding settings, and subset–round seed schedule are fixed independently of the candidate menu. Each subset therefore has a fixed response for a fixed seed table.

Let $F(S) = a$ when at least four of the five R2 answers agree on normalized answer a, and $F(S) = \perp$ otherwise. A nonempty set with $F(S) = a$ is a *witness* for a. Invalid outputs and abstentions remain in the five-vote denominator; the parser accepts complete JSON answers without repair, and `unknown` maps to $\perp$.

Each replay also reveals $D(S)$, the valid non-abstaining answers produced by any agent at any round. After querying Q_t, the candidate menu is $C_t = \bigcup_{S \in Q_t} D(S)$, initialized with $Q_0 = \{V, \varnothing\}$. Let $P_{a,t}$ be candidate a's observed nonempty witnesses, and N_t the unqueried nonempty subsets. A candidate can enter the menu without final consensus; it gains a witness when $F(S) = a$. Later queries can discover answers absent from the initial menu.

For a family $\mathcal{F}$ of nonempty evidence subsets, a *packing* is a collection of pairwise disjoint members. We write $\nu(\mathcal{F})$ for its maximum cardinality, with $\nu(\varnothing) = 0$. This operator counts subsets that can coexist without reusing evidence.

3.2 A CAPACITY OBJECTIVE FOR ANSWER SUPPORT

A supporting set must retain the passages that work together, while its passages should not be counted again through overlapping sets. We impose one unit of capacity on each evidence unit and maximize the number of complete witnesses retained. If $F(\varnothing) = a$, we set $K(a) = 0$: support reproduced without evidence contributes no evidence-derived capacity. For every other answer, let $\mathcal{W}_a = \{S \subseteq V : S \neq \varnothing,\ F(S) = a\}$ and define

$$K(a) = \max_{x \in \{0,1\}^{\mathcal{W}_a}} \sum_{S \in \mathcal{W}_a} x_S \quad \text{subject to} \quad \sum_{\substack{S \in \mathcal{W}_a \\ u \in S}} x_S \leq 1 \quad \text{for every } u \in V. \tag{1}$$

Selecting $x_S = 1$ retains the complete witness S. Thus $\{u_1, u_2\}$ can count even when neither singleton supports the answer. It cannot count alongside $\{u_1, u_3\}$, but can coexist with $\{u_3, u_4\}$. The objective preserves complementarity within each witness and prevents reuse across the selected witnesses.

The binary constraint makes capacity a count of complete supports. For example, the pairwise-overlapping witnesses $\{u_1, u_2\}$, $\{u_2, u_3\}$, and $\{u_1, u_3\}$ have capacity one. Giving each weight one half would satisfy the unit constraints but would not select a disjoint collection. The same unit constraints bound the influence of local evidence changes: modifying k unit positions changes capacity by at most k when replays avoiding those positions remain fixed (Proposition 1).

3.3 CAPACITY BOUNDS FROM INCOMPLETE OBSERVATIONS

The observed witnesses $P_{a,t}$ reveal only part of the full witness family. For an answer not supported by the empty control, we bound its capacity using the known witnesses and the unqueried subsets:

$$L_t(a) = \nu(P_{a,t}) \ \leq \ K(a) \ \leq \ H_t(a) = \nu(P_{a,t} \cup N_t). \tag{2}$$

Known witnesses give the lower bound; allowing every unqueried set to support a gives the upper bound. Both bounds are zero for an empty-supported answer. A failed query rules out only that subset, since adding or removing evidence can change the response.

An undiscovered answer has no observed witness, so its capacity is bounded by

$$K(a) \leq H_t^{\text{new}} = \nu(N_t) \qquad \text{for } a \notin C_t. \tag{3}$$

This bound tracks support still available outside the discovered menu and identifies when the entire capacity state is resolved.

3.4 Deriving the optimal one-step query

A replay can strengthen an answer or rule out a possible support. Its value therefore depends on how both outcomes change the capacity intervals. We measure uncertainty by their total width, $W_t = \sum_{a \in C_t} [H_t(a) - L_t(a)]$. Let $W_t(S, y)$ be that width after querying S and observing response y, with the comparison taken over the current menu C_t. The acquisition objective is

$$S_{t+1} \in \arg\min_{S \in N_t} \max_{y \in C_t \cup \{\star\}} W_t(S, y), \tag{4}$$

where $\star$ groups no consensus and answers outside C_t. The menu is fixed for this one-step comparison; all answers in $D(S)$ enter the next state. Each replay consumes one query unit.

How a response changes the bounds. For an answer not supported by the empty control and $S \in N_t$, define its positive and negative response sensitivities:

$$\begin{aligned} \alpha_{a,t}(S) &= \nu(P_{a,t} \cup \{S\}) - L_t(a), \\ \beta_{a,t}(S) &= H_t(a) - \nu\big(P_{a,t} \cup (N_t \setminus \{S\})\big). \end{aligned} \tag{5}$$

Set both to zero for an empty-supported answer. Each sensitivity is zero or one: inserting or deleting one witness changes the maximum packing by at most one. A response supporting a increases its lower bound by $\alpha_{a,t}(S)$ and leaves its upper bound fixed. Any other response reduces its upper bound by $\beta_{a,t}(S)$ and leaves its lower bound fixed.

The negative sensitivity distinguishes a necessary member of every maximum optimistic packing from a member of just one such packing. Querying the former can eliminate a possible unit of capacity. The positive sensitivity asks whether the same set can add a complete disjoint witness. These two effects yield an explicit minimax score.

Theorem 1 (Optimal one-step acquisition)**.** *Under the fixed replay model, the smallest reduction in current-menu interval width guaranteed by querying S is*

$$G_t(S) = \sum_{a \in C_t} \beta_{a,t}(S) - \mathbf{1}\{\exists a \in C_t : \beta_{a,t}(S) = 1,\ \alpha_{a,t}(S) = 0\}\,. \tag{6}$$

Hence any $S_{t+1} \in \arg\max_{S \in N_t} G_t(S)$ solves Eq.(4)*.*

To see why, a response $y \in C_t$ reduces total width by $\sum_a \beta_{a,t}(S) + \alpha_{y,t}(S) - \beta_{y,t}(S)$; an outside-menu response reduces it by $\sum_a \beta_{a,t}(S)$. Taking the smallest of these reductions gives $G_t(S)$. Appendix A.3 supplies the complete proof. CWE evaluates this score for every unqueried subset and breaks ties by the fixed subset order in Appendix B.1.

Reading the score. Consider a query with $\alpha_a = \alpha_b = 1$, $\beta_a = 0$, and $\beta_b = 1$. If it supports a, the lower bound for a rises and the upper bound for b falls, reducing total width by two. If it supports b, that answer's lower bound rises. Any other response lowers the upper bound for b. The query therefore guarantees a reduction of one, exactly its score $G = 1$. The criterion rewards a query whose possible responses resolve capacity uncertainty in different ways.

Witness exchange. Each new observation triggers a fresh packing computation. If the selected witness is $\{u_1, u_2\}$ and later replays reveal $\{u_1, u_3\}$ and $\{u_2, u_4\}$ for the same answer, neither can be appended to the old selection. Together they form a larger disjoint collection. Exchanging the old witness for these two raises the lower bound from one to two. Both sensitivities in Eq.(5) recompute packings over the full family, so occupied evidence remains available to such exchanges.

3.5 Stopping and answer selection

CWE certifies a unique capacity winner among the discovered candidates when

$$L_t(a) > \max_{b \in C_t \setminus \{a\}} H_t(b). \tag{7}$$

The maximum over an empty set is zero, so a certificate requires positive support. The capacity bounds then imply $K(a) > K(b)$ for all other $b \in C_t$. This *closed-candidate certificate* compares the discovered menu. Search also stops when every observed interval is exact and $H_t^{\text{new}} = 0$, or when the additional-query allowance is exhausted. At any stop, CWE selects a candidate maximizing $L_t(a)$ under a fixed tie order; an empty menu returns $\perp$. Algorithm 1 summarizes the procedure.

Computation and replay cost. Packing and sensitivity evaluation use exact subset enumeration, exponential in the number of evidence units. The five-unit contexts evaluated here have 32 subsets. Computation of G_t uses the current observations and does not require additional model responses. A complete replay produces 15 agent outputs across R0–R2. The full and empty controls both incur this cost; the Report convention counts them jointly as one initialization unit, while token totals include both.

Algorithm 1 Counterfactual Witness Exchange

Require: Canonical evidence V, fixed replay rule, additional-query allowance B
Replay V and $\varnothing$; set $Q \leftarrow \{V, \varnothing\}$ and $C \leftarrow D(V) \cup D(\varnothing)$
Compute $L(a), H(a)$ for $a \in C$ and H^{new} by Eqs.(2)–(3)
for $j = 1, \ldots, B$ **do**
  if a discovered candidate satisfies Eq.(7) **then**
    break with a closed-candidate certificate
  end if
  if $H^{\text{new}} = 0$ and $L(a) = H(a)$ for every $a \in C$ **then**
    break with an exact capacity state
  end if
  Compute $\alpha_a(S), \beta_a(S)$ and $G(S)$ for every unqueried nonempty S
  Choose $S \in \arg\max G(S)$, breaking ties by the fixed subset order
  Replay S; update $Q \leftarrow Q \cup \{S\}$ and $C \leftarrow C \cup D(S)$
  Recompute all capacity bounds and the candidate tie order
end for
Check Eq.(7) on the final bounds for certification
return a candidate maximizing $L(a)$ under the fixed tie rule, or $\perp$ if $C = \varnothing$

4 Experiments

4.1 Evaluation setup

We evaluate answer quality and realized cost under a common query allowance, then use fixed-observation and structural comparisons to explain the results.

Data and models. We sample 500 questions from the 7,405-question HotpotQA distractor development split (Yang et al., 2018) and 500 from the 12,576-question 2WikiMultiHopQA development split (Ho et al., 2020). The backbones are Qwen2.5-7B, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3, with seeds 42, 3407, and 1337. Each method has 9,000 question–model–seed records. Replays use five agents at R0–R2, temperature 0.2, and top-$p = 0.95$. All methods receive the same questions, canonical evidence, and model settings.

Comparison policies. The subset policies share candidate discovery, capacity aggregation, and stopping rules. Each policy sees an answer only after acquiring the replay that produces it. The comparison therefore changes how evidence is queried while retaining a common answer-selection procedure.

Singleton-first queries sets in increasing size, starting with individual units. Leave-one-out-first (LOO-first) reverses this order: it first removes one unit from the full context, then tests progressively smaller sets. These schedules approach complementary evidence from opposite directions. Static grouping precommits cyclic partitions before observing responses. It starts with pair partitions, including any leftover singleton, then uses singleton and larger-group passes. Fixed-order queries complete the schedule. Uniform random permutes the unqueried subsets; scheduler seeds are averaged within each record.

Greedy adapts its queries to observed support. It finds a witness for an answer and repeatedly deletes an evidence unit whenever the smaller set still supports that answer. It then commits the remaining units and searches unused evidence for further witnesses. A fixed-order fallback fills remaining queries and can revisit occupied units. CWE selects each query by its guaranteed one-step reduction in capacity uncertainty; all subset policies retain exact packing for answer selection. The dedicated query restriction in Section 4.3 tests this exchange mechanism directly.

Repeated full-context inference votes over eight full-context generations. A single trajectory supplies the low-cost endpoint. These controls measure the return from spending model calls on the unchanged context, alongside the policies that vary the evidence available to each replay. Appendix B.1 gives the deterministic tie and subset orders.

Cost and uncertainty. The allowance is eight *Reports*. For subset policies, full and empty initialization jointly count as one Report, and each additional replay adds one. Repeated inference counts each full-context generation as one and uses no empty control. A subset replay contains 15 internal agent outputs. Tokens include all inputs and outputs, including initialization. We report EM, token F1, and cost per record, retaining invalid and unanswered cases in the denominator. EM requires equality after answer normalization; F1 is the harmonic mean of token precision and recall against the reference. Mismatches involving `yes`, `no`, or `noanswer` receive zero F1. Normalization lowercases text, removes punctuation and articles, and standardizes whitespace. Paired differences are averaged over the fixed models and seeds within each question before dataset-stratified case resampling. The 95% intervals describe variation over questions.

4.2 Answer quality and inference cost

Table 1 summarizes the main comparison. CWE reaches .53 EM and .65 F1. Relative to greedy, both gains are .03 $[.00, .06]$, with a token difference of -0.4k $[-0.8, 0.0]$k. The intervals include zero. Relative to repeated full-context inference, both gains are .08 $[.05, .11]$, with 11.8k fewer tokens $[11.3, 12.3]$k. The single trajectory remains the lowest-cost option. CWE has the highest quality point estimates in all six dataset–backbone cells, with a .03 EM advantage over greedy in each (Appendix B).

Table 1: Answer quality and cost across 9,000 records per method. EM and F1 are fractions; tokens include every input and output.

Method	EM ↑	F1 ↑	Reports	Tokens
Single full-context trajectory	.44	.56	1.0	3.0k
Repeated full-context inference	.45	.57	8.0	24.5k
Singleton-first	.48	.60	6.2	13.0k
Static grouping	.47	.59	5.0	12.5k
LOO-first	.49	.61	6.8	14.2k
Greedy	.50	.62	5.5	13.1k
Uniform random	.45	.57	5.0	12.4k
CWE	**.53**	**.65**	5.4	12.7k

At allowances two, four, and eight, CWE reaches .45, .47, and .53 EM, compared with .46, .48, and .50 for greedy. At allowance four, both use 3.9 Reports and approximately 8.3k versus 8.5k tokens. Repeated full-context inference rises from .44 to .45 EM across these allowances, while token use grows from 6.2k to 24.5k (Table 4).

The allowance-four comparison isolates query choices at a similar realized cost. At allowance eight, early stopping and shorter subset prompts also affect cost, which is why we report both replay counts and tokens.

An exhaustive reference on 200 questions evaluates all 32 subsets of five evidence units. With $K^* = \max_a K(a)$, the observed-capacity gap $G_{\text{obs}} = K^* - \max_a L_t(a)$ measures undiscovered support; the selection gap $G_{\text{sel}} = K^* - K(\hat{a}_t)$ measures the selected answer's capacity shortfall. At allowance eight, CWE's gaps are .4 and .3, compared with .6 and .4 for greedy (Table 4). Selection by maximum lower bound ensures $0 \le G_{\text{sel}} \le G_{\text{obs}}$.

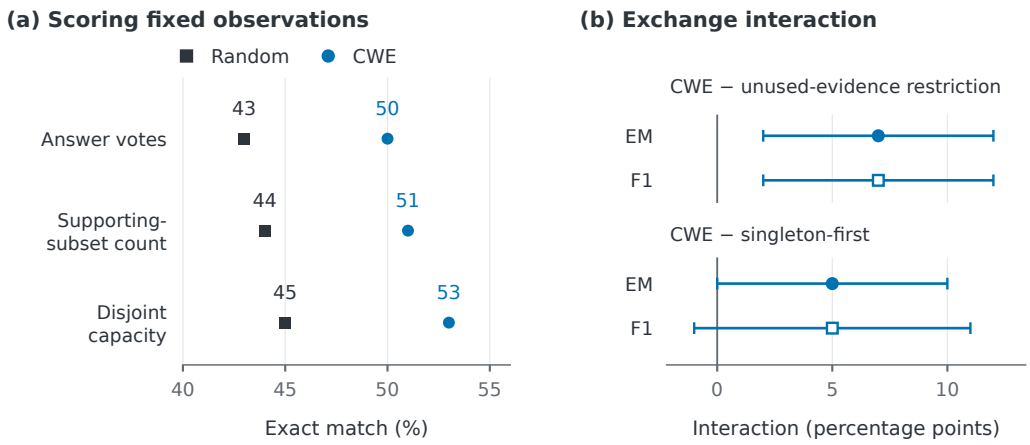

Figure 2: **Capacity scoring and witness exchange.** (a) Answer EM after rescoring the same observations within each query policy. (b) Additional CWE advantage when support requires an exchange, relative to non-obstructing structures. Bars are paired 95% intervals.

4.3 What contributes to the gain?

Figure 2 separates answer aggregation from query selection. Holding the observations fixed, capacity scoring improves EM over answer votes by .03 on CWE queries and .02 on random queries. With capacity scoring held fixed, changing from random to CWE queries raises EM by .08, with realized costs of 5.0 versus 5.4 Reports and 12.4k versus 12.7k tokens (Table 5).

Accounting for overlap under fixed observations. Votes count answer occurrences; supporting-subset counts record how often a subset reproduces an answer; capacity also uses the overlap between those subsets. On a fixed prefix, the queries, candidate menu, and replay outcomes are identical. Capacity improves EM over subset counts from .51 to .53 on CWE prefixes and from .44 to .45 on random prefixes. These comparisons isolate the value of accounting for reuse in final selection. The next experiment holds this scoring rule fixed and changes which subsets can be queried.

The query-restriction experiment tests the role of exchange directly. A restricted policy commits an initial witness and subsequently queries only unused evidence. All policies receive the same paid prefix and use exact capacity aggregation. The comparison includes 120 matched questions per condition, three initial-witness variants, and two identifier/order variants. CWE gains .09 EM over the restriction when the initial witness blocks alternatives, versus .02 when it does not. The paired interaction is .07 $[.02, .12]$ for both EM and F1. Against singleton-first, the EM interaction is .05 $[.00, .10]$, while the F1 interval spans zero (Table 7).

Queries also expand the candidate menu. Correct-answer coverage rises from .61 initially to .80 under CWE, compared with .77 for greedy and .74 for singleton-first. Newly discovered correct answers are selected in .09 of all CWE records. On the same CWE observations, selecting from the final menu instead of the initial menu raises EM from .50 to .53 (Table 5).

A newly encountered answer can enter the menu before any subset reproduces it with four final votes. Later queries test its support. Candidate coverage therefore measures the opportunity for a correct selection, while EM measures the choice made. The final-menu comparison holds queries and support scores fixed, isolating the net effect of allowing newly discovered answers to be selected.

4.4 SUPPORT STRUCTURE AND OBSERVED LIMITS

The structural experiment uses 150 cases per condition: a shared necessary bridge, disjoint alternative supports, and a missing necessary hop. Replacing the shared bridge with alternatives raises correct-answer capacity from 1.1 to 2.4, a paired increase of 1.3 $[1.0, 1.6]$. The capacity margin over the strongest wrong answer rises by 1.1 $[0.8, 1.4]$, alongside an EM gain of .17 $[.11, .23]$. Removing a necessary hop instead lowers capacity by .5 $[.3, .7]$ and EM by .12 $[.06, .18]$. Initial agreement differs across these interventions (.78, .52, .66). Table 6 reports the full condition profiles and realization checks.

Why the structural intervention is informative. The condition ordering changes when support is measured beyond the complete context. Disjoint alternatives have lower initial agreement than the shared bridge, yet higher correct-answer capacity, a larger capacity margin, and higher CWE accuracy. This pattern matches the proposed role of alternative support: several compatible routes to an answer can be useful even when the full-context agents agree less. Removing a necessary hop lowers both capacity and accuracy, connecting the score to whether a complete supporting set remains available. These interventions connect the capacity score to both parts of the support structure: combining passages within a witness and retaining several witnesses without reusing a passage.

High capacity can still favor a wrong answer. CWE certifies 4,500 of 9,000 records, of which 675 are wrong, giving a conditional error rate of .15. In the exhaustive subset, 117 of 1,800 selections are wrong and have greater capacity than the correct answer. The empty-control rule also suppresses useful priors: rescoring the same queries without it corrects 410 answers and loses 250, a net EM gain of .018 $[.008, .028]$. Among 300 reviewed errors, 96 involve missing necessary evidence and 27 involve normalization errors (Table 8). These outcomes distinguish the support recovered by the model from answer correctness.

5 CONCLUSION

CWE uses the maximum number of complete supporting sets that can coexist without reusing evidence as its objective. This preserves multi-hop complementarity and bounds the influence of local evidence changes on the support score. A second objective minimizes worst-case capacity uncertainty at each query. Its exact one-step solution follows from the positive and negative sensitivities of packing bounds. Witness exchange allows newly discovered alternatives to replace an earlier support. The experiments connect answer quality to both overlap-aware scoring and the discovery of alternative supports, with the clearest exchange effect against a policy restricted to unused evidence.

The current implementation uses exact search over five-unit contexts, with exponential cost as the number of units grows. Disjoint texts may share an upstream source or elicit a common reasoning error. Together with the observed cost of empty-control zeroing, these results identify the next design problem: combining measured external support with useful model priors.

Reproducibility statement. Appendix B provides search details, dataset breakdowns, paired contrasts, ablations, and failure results. The accompanying diagnostic supplement details the communication and matched-lineage experiments behind the motivation, including their protocols, case-paired analyses, and identifier and team controls.

REFERENCES

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, 1992.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei A. Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Justin 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 (Volume 1: Long Papers)*, pp. 7066–7085, 2024.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Mądry. ContextCite: Attributing model generation to context. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95764–95807, 2024.

Yu Cui, Hang Fu, Haibin Zhang, Licheng Wang, and Cong Zuo. Free-mad: Consensus-free multi-agent debate. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31977–31997, 2026.

Peter M. DeMarzo, Dimitri Vayanos, and Jeffrey Zwiebel. Persuasion bias, social influence, and unidimensional opinions. *The Quarterly Journal of Economics*, 118(3):909–968, 2003.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020.

Xin Luna Dong, Laure Berti-Équille, and Divesh Srivastava. Integrating conflicting data: The role of source dependence. *Proceedings of the VLDB Endowment*, 2(1):550–561, 2009.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763, 2024.

Chen Han, Wenzhen Zheng, and Xijin Tang. Debate-to-detect: Reformulating misinformation detection as a real-world debate with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15114–15129, 2025.

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, 2020.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 30038–30066, 2025.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

Hithesh Sankararaman, Mohammed Nasheed Yasin, Tanner Sorensen, Alessandro Di Bari, and Andreas Stolcke. Provenance: A light-weight fact-checker for retrieval augmented LLM generation output. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1305–1313, 2024.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024.

Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D Barrett, and Arnu Pretorius. Should we be going MAD? a look at multi-agent debate strategies for LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 45883–45905, 2024.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 10131–10165, 2025.

Tanel Tammet, Dirk Draheim, and Priit Järv. Confidences for commonsense reasoning. In André Platzer and Geoff Sutcliffe (eds.), *Automated Deduction – CADE 28*, volume 12699 of *Lecture Notes in Computer Science*, pp. 507–524, 2021.

Hongduan Tian, Xiao Feng, Ziyuan Zhao, Xiangyu Zhu, Rolan Yan, and Bo Han. Multi-agent debate with memory masking. In *International Conference on Learning Representations*, 2026.

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.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Kevin Wu, Eric Wu, and James Zou. ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust RAG against retrieval corruption. *arXiv preprint arXiv:2405.15556v2*, 2026.

Ruoyu Xu, Gaoxiang Li, and Victor S. Sheng. GAVEL: Evidence-contract debate with mechanized scrutiny for provenance-grounded fact-checking. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35907–35920, 2026.

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.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 11809–11822, 2023.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14544–14607, 2024.

Siyuan Zhu, Chengdong Xu, Kaiqiang Ke, and Chao Yu. Context-Picker: Dynamic context selection using multi-stage reinforcement learning. *arXiv preprint arXiv:2512.14465v2*, 2026a.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. Conformity in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3854–3872, 2025.

Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. Demystifying multi-agent debate: The role of confidence and diversity. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 33909–33930, 2026b.

AI USE STATEMENT

Generative AI tools were used only for recommended tasks: polishing manuscript wording, improving readability, and formatting reference entries. No generative AI was used to generate synthetic datasets, develop theoretical frameworks, formulate hypotheses, design experiments, implement core methods, interpret experimental results, or prove mathematical claims. All AI-assisted text was manually reviewed, revised, and verified by the authors. The authors take full responsibility for all claims, arguments, and content of this work.

A SUPPORT CAPACITY AND ACQUISITION GUARANTEES

The results below explain two properties of CWE. Disjoint witnesses limit how much a local evidence change can alter an answer's capacity. The response sensitivities then yield the exact query that minimizes the worst-case interval width after one replay.

A.1 HOW EVIDENCE CHANGES AFFECT CAPACITY

Compare two replay functions F and F' on the same canonical unit positions. Let $J \subseteq V$ contain the modified positions. A replay whose subset avoids J receives the same evidence in the same order, with the same prompt and seed schedule. The unit indexing stays fixed, including when replacement texts would otherwise change preprocessing.

Proposition 1 (Bounded evidence influence)**.** *If $F'(S) = F(S)$ whenever $S \cap J = \varnothing$, then for every non-abstaining answer a,*

$$|K_{F'}(a) - K_F(a)| \leq |J|. \tag{8}$$

A changed position can affect many witnesses, but it can belong to at most one witness in a disjoint packing. This gives the bound.

Proof of Proposition 1. Fix a non-abstaining answer a. The assumption also covers the empty subset, so $F'(\varnothing) = F(\varnothing)$. If $F(\varnothing) = a$, the empty-control rule sets both capacities to zero.

Otherwise, take a maximum witness packing $\mathcal{P}$ for a under F and retain the witnesses that avoid the modified positions:

$$\mathcal{P}_0 = \{S \in \mathcal{P} : S \cap J = \varnothing\}.$$

Every removed witness contains a position in J. These positions are distinct because the witnesses are disjoint, so at most $|J|$ witnesses are removed. Every retained witness still satisfies $F'(S) = F(S) = a$. Consequently,

$$K_{F'}(a) \geq |\mathcal{P}_0| \geq K_F(a) - |J|.$$

Exchanging F and F' gives $K_F(a) \geq K_{F'}(a) - |J|$ and proves the absolute bound. □

A.2 WHEN THE FULL-CAPACITY WINNER STAYS UNCHANGED

Changing at most k unit positions can lower the winner's capacity by at most k and raise a competitor's by at most k. A margin greater than $2k$ therefore preserves the winner.

Corollary 1. *Under the assumptions of Proposition 1, let $|J| \leq k$. If an answer a satisfies*

$$K_F(a) - \max_{b \neq a} K_F(b) > 2k, \tag{9}$$

then a is the unique capacity winner under F'. The maximum ranges over all other non-abstaining normalized answers, including answers not observed under F, and is zero if there are none.

Proof. For every competing answer b, Proposition 1 implies

$$K_{F'}(a) - K_{F'}(b) \geq K_F(a) - K_F(b) - 2k > 0.$$

Newly appearing answers are included because their original capacity is zero. □

The corollary concerns capacities over all evidence subsets. The same packing argument applies to scores on a common fixed query family. A budgeted adaptive run can change its queried family when the evidence changes, so its output is not covered by this result. The stopping rule in Eq.(7) instead certifies a capacity winner among the candidates discovered in that run.

A.3 Deriving the one-step minimax query

Proof of Theorem 1. Fix the observations at round t and a nonempty query family N_t. The objective measures width over the current candidate menu C_t: $W_t = \sum_{a \in C_t} [H_t(a) - L_t(a)]$. Write $W_t(S, y)$ for this width after querying S and receiving response y. New answers enter C_{t+1} through Eq.(14); they do not change the menu used in this one-step comparison. All queries have unit replay cost.

Step 1: Each sensitivity is zero or one. Inserting one member into a set family cannot decrease its packing number. It can increase the number by at most one: removing that member from a new packing leaves a feasible packing in the old family. Reversing the argument gives the corresponding deletion bound. Thus $\alpha_{a,t}(S), \beta_{a,t}(S) \in \{0, 1\}$. For an empty-control-supported candidate, both are zero because its capacity bounds remain zero.

For any other candidate, $\beta_{a,t}(S) = 1$ exactly when every maximum optimistic packing contains S. An optimum avoiding S survives its deletion. If every optimum contains S, deleting it lowers the optimum by one. This is why negative sensitivity measures whether the query can remove a whole unit of possible capacity.

Step 2: Write the bounds after each response. For a candidate a not supported by the empty control, response $y = a$ moves S from N_t to $P_{a,t}$. Their union stays fixed, so the upper bound is unchanged and the lower bound rises by $\alpha_{a,t}(S)$. Response $y \neq a$ leaves the known witnesses unchanged and removes S from the optimistic family, lowering the upper bound by $\beta_{a,t}(S)$. Including the empty-control case, the exact updates are

$$\begin{aligned} L_t^+(a; S, y) &= L_t(a) + \mathbf{1}\{y = a\}\alpha_{a,t}(S), \\ H_t^+(a; S, y) &= H_t(a) - \mathbf{1}\{y \neq a\}\beta_{a,t}(S). \end{aligned} \tag{10}$$

These updates apply to every interval in the objective, regardless of which answer the replay supports.

Step 3: Sum the width reduction. Let $B_t(S) = \sum_{a \in C_t} \beta_{a,t}(S)$. Summing the bound updates gives

$$W_t - W_t(S, y) = \begin{cases} B_t(S) + \alpha_{y,t}(S) - \beta_{y,t}(S), & y \in C_t, \\ B_t(S), & y \notin C_t \text{ or } y = \perp. \end{cases} \tag{11}$$

A response outside the current menu, including a newly supported answer, is a negative observation for every current candidate. Discovering additional candidates through $D(S)$ also leaves the bounds of current candidates unchanged.

Step 4: Take the worst case and choose the query. Every branch in Eq.(11) is compatible with the observations: assign that response to the unqueried subset S and retain all observed responses. The replay model imposes no relation between responses on different subsets. The smallest possible decrease is therefore

$$\begin{aligned} \min_y \{W_t - W_t(S, y)\} &= B_t(S) + \min\big(\{0\} \cup \{\alpha_{a,t}(S) - \beta_{a,t}(S) : a \in C_t\}\big) \\ &= B_t(S) - \mathbf{1}\{\exists a \in C_t : \beta_{a,t}(S) = 1,\ \alpha_{a,t}(S) = 0\} \\ &= G_t(S). \end{aligned} \tag{12}$$

The set form includes an empty candidate menu. Since W_t is independent of the chosen query, $\max_y W_t(S, y) = W_t - G_t(S)$. Maximizing $G_t(S)$ therefore minimizes the largest possible remaining width after one replay. The fixed tie rule selects one of these minimizers. □

B Additional CWE results

B.1 Deterministic search and packing details

Candidate discovery. For agent i and round r, let $a_{i,r}(S)$ be the normalized answer on subset S, with invalid output represented by $\perp$. The candidate set revealed by one replay is

$$D(S) = \{a_{i,r}(S) : 1 \leq i \leq 5,\ 0 \leq r \leq 2\} \setminus \{\perp, \texttt{unknown}\}. \tag{13}$$

Algorithm 1 adds observed labels to the candidate menu and updates support using $F(S)$. An answer can therefore become a candidate before it receives enough final votes to become the support response.

Candidate and subset orders. Within each initialization trajectory, answers are ordered by first appearance round, descending votes in that round, then answer text. The initial menu lists $D(V)$ before new members of $D(\varnothing)$. Later discoveries form a lexically sorted suffix, making the final order independent of query order after complete enumeration. Subsets are ordered by size, then canonical unit indices.

Computing the acquisition score. CWE computes Eq.(5) for each current candidate and unqueried nonempty subset. Empty-supported candidates have zero sensitivities. It evaluates $G_t(S)$ by Eq.(6) and queries the first maximizer in subset order, including when all scores are zero or the menu is empty. Packing ties follow the same order; acquisition depends only on optimum cardinalities.

After receiving $F(S)$ and $D(S)$, the state update is

$$\begin{aligned} C_{t+1} &= C_t \cup D(S), & N_{t+1} &= N_t \setminus \{S\}, \\ P_{a,t+1} &= P_{a,t} \cup \{S : F(S) = a\}, & a &\in C_{t+1}, \end{aligned} \tag{14}$$

where $P_{a,t} = \varnothing$ for a newly discovered answer. All capacity bounds are recomputed before stopping or choosing the next query.

Replay cost and evidence units. For A agents and R communication rounds, a complete replay produces $A(R+1)$ agent outputs: 15 for five agents at R0–R2. In the main panel, full and empty initialization jointly count as one Report, and B subsequent queries give $1 + B$ Reports. Token costs include both controls and every subsequent replay. The earlier development tables count individual agent outputs.

Canonicalization removes exact-text duplicates and ignores identifier names. With canonical prompts and replay randomness fixed, duplicating a text or renaming its identifier leaves CWE unchanged. Different texts remain separate units even when their origins overlap.

B.2 Performance by dataset and backbone

Table 2 separates results by dataset and backbone. Each cell contains 500 questions evaluated with seeds 42, 3407, and 1337, giving 1,500 question–seed records. Qwen, Llama, and Mistral denote Qwen2.5-7B, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3. CWE has the highest EM and F1 in all six cells. Reports use the convention above; Tokens include both initialization controls. Rounding explains small cost differences from Table 1.

Table 2: Answer quality and cost by backbone. Each dataset–backbone cell averages 1,500 question–seed records. EM and F1 are fractions. The two datasets have identical reported cost means at the shown precision, so Reports and Tokens share columns; Tokens are in thousands.

	HotpotQA		2WikiMultiHopQA			
Method	EM	F1	EM	F1	Reports	Tokens (k)
Qwen2.5-7B						
Single full context	0.420	0.540	0.450	0.570	1.0	3.0
Repeated full context	0.430	0.550	0.460	0.580	8.0	24.0
Singleton-first	0.460	0.580	0.490	0.610	6.1	12.7
Static grouping	0.450	0.570	0.480	0.600	5.0	12.2
LOO-first	0.470	0.590	0.500	0.620	6.7	13.9
Greedy	0.480	0.600	0.510	0.630	5.4	12.8
Uniform random	0.430	0.550	0.460	0.580	5.0	12.1
CWE	0.510	0.630	0.540	0.660	5.3	12.4

Table 2 (continued)

	HotpotQA		2WikiMultiHopQA			
Method	EM	F1	EM	F1	Reports	Tokens (k)
Llama-3.1-8B-Instruct						
Single full context	0.455	0.575	0.485	0.605	1.0	3.2
Repeated full context	0.465	0.585	0.495	0.615	8.0	26.0
Singleton-first	0.495	0.615	0.525	0.645	6.4	13.8
Static grouping	0.485	0.605	0.515	0.635	5.1	13.3
LOO-first	0.505	0.625	0.535	0.655	7.0	15.1
Greedy	0.515	0.635	0.545	0.665	5.7	13.9
Uniform random	0.465	0.585	0.495	0.615	5.1	13.1
CWE	0.545	0.665	0.575	0.695	5.6	13.5
Mistral-7B-Instruct-v0.3						
Single full context	0.400	0.520	0.430	0.550	1.0	2.9
Repeated full context	0.410	0.530	0.440	0.560	8.0	23.5
Singleton-first	0.440	0.560	0.470	0.590	6.1	12.5
Static grouping	0.430	0.550	0.460	0.580	4.9	12.0
LOO-first	0.450	0.570	0.480	0.600	6.7	13.6
Greedy	0.460	0.580	0.490	0.610	5.4	12.6
Uniform random	0.410	0.530	0.440	0.560	4.9	11.9
CWE	0.490	0.610	0.520	0.640	5.3	12.2

B.3 Paired quality and cost differences

Table 3 reports CWE minus each comparator. Positive quality differences favor CWE; negative cost differences indicate savings. Pairing is retained within question, backbone, and seed, with uncertainty estimated over questions. EM and F1 share a column because their reported estimates and intervals coincide. CWE improves quality and reduces cost relative to repeated full-context inference. Against Greedy, the quality and cost intervals include or touch zero. Dashes indicate unavailable paired cost estimates.

Table 3: Paired differences in quality and cost, measured as CWE minus each comparator. Brackets give 95% intervals. EM and F1 use the 0–1 scale; token differences are in thousands.

Comparator	ΔEM and ΔF1	ΔReports	ΔTokens (k)
Single full context	$+0.090$ $[+0.060, +0.120]$	—	—
Repeated full context	$+0.080$ $[+0.050, +0.110]$	-2.6 $[-2.7, -2.5]$	-11.8 $[-12.3, -11.3]$
Singleton-first	$+0.050$ $[+0.020, +0.080]$	-0.8 $[-1.1, -0.5]$	-0.3 $[-0.7, +0.1]$
Static grouping	$+0.060$ $[+0.030, +0.090]$	—	—
LOO-first	$+0.040$ $[+0.010, +0.070]$	—	—
Greedy	$+0.030$ $[+0.000, +0.060]$	-0.1 $[-0.4, +0.2]$	-0.4 $[-0.8, +0.0]$
Uniform random	$+0.080$ $[+0.050, +0.110]$	—	—

B.4 Query budget and search completeness

Table 4 links the query allowance to answer quality and recovered support. Panel A uses the main panel. Allowances include initialization; $\approx$ marks approximate token costs. Panel B uses an exhaustive reference on the first 100 sampled questions from each dataset: 200 questions and 1,800 records across three backbones and three seeds. All 32 subsets are replayed offline; their cost is excluded from online Reports and Tokens.

With $K^* = \max_a K(a)$, the observed-support gap $G_{\text{obs}} = K^* - \max_a L_t(a)$ measures unrecovered capacity. The selection gap $G_{\text{sel}} = K^* - K(\hat{a}_t)$ measures the chosen answer's shortfall from the best exact capacity. Lower gaps are better. Cert. is the fraction with a certified unique capacity winner among discovered candidates. Increasing the allowance improves CWE's quality and reduces both gaps.

Table 4: Effect of the query allowance. Panel A reports main-panel quality and realized cost. Panel B measures remaining capacity gaps and certification against the exhaustive reference over 1,800 records. Certification covers discovered candidates.

A. Quality and realized cost

Allowance	Policy	EM	F1	Reports	Tokens
2	Repeated full context	0.44	0.56	2.0	6.2k
2	Singleton-first	0.45	0.57	2.0	4.3k
2	Greedy	0.46	0.58	2.0	4.4k
2	CWE	0.45	0.57	2.0	4.3k
4	Repeated full context	0.45	0.57	4.0	12.3k
4	Singleton-first	0.47	0.59	3.9	$\approx$ 8.3k
4	Greedy	0.48	0.61	3.9	$\approx$ 8.5k
4	CWE	0.47	0.59	3.9	$\approx$ 8.3k
8	Repeated full context	0.45	0.57	8.0	24.5k
8	Singleton-first	0.48	0.60	6.2	13.0k
8	Greedy	0.50	0.62	5.5	13.1k
8	CWE	0.53	0.65	5.4	12.7k

B. Capacity gaps and certification

Allowance	Policy	G_{obs}	G_{sel}	Cert.	EM
2	Singleton-first	1.6	1.4	0.00	0.45
2	Greedy	1.5	1.3	0.00	0.46
2	CWE	1.6	1.4	0.00	0.45
4	Singleton-first	1.1	0.9	0.06	0.47
4	Greedy	1.0	0.8	0.09	0.48
4	CWE	1.1	0.9	0.06	0.47
8	Singleton-first	0.7	0.5	0.38	0.48
8	Greedy	0.6	0.4	0.44	0.50
8	CWE	0.4	0.3	0.52	0.53

B.5 Answer scoring and candidate discovery

Table 5 separates answer scoring from candidate discovery. Panel A applies three answer rules to each saved query sequence up to stopping. Candidates, trajectories, empty-control treatment, tie order, and cost stay fixed within a sequence. Disjoint capacity gives the highest EM and F1 for both random and CWE queries.

Panel B tracks discovery from $C_0 = D(V) \cup D(\varnothing)$ to the final menu. Coverage is the fraction of all records containing an EM-correct candidate. New-correct is the fraction whose final answer is correct

and absent from C_0. Panel C reranks the same saved CWE queries using the initial or final menu. Capacity scores, tie order, and costs stay fixed; online candidate discovery is unchanged. Allowing the final menu at selection raises EM from .50 to .53.

Table 5: Answer scoring and candidate discovery at the main allowance. Panel A changes the scoring rule on fixed queries; Panel B tracks candidate coverage; Panel C changes the selectable menu on fixed CWE queries.

A. Answer aggregation on saved prefixes

Prefix source	Answer rule	EM	F1	Reports	Tokens
Random	Answer votes	0.43	0.55	5.0	12.4k
Random	Supporting-subset count	0.44	0.56	5.0	12.4k
Random	Disjoint capacity	0.45	0.57	5.0	12.4k
CWE	Answer votes	0.50	0.62	5.4	12.7k
CWE	Supporting-subset count	0.51	0.63	5.4	12.7k
CWE	Disjoint capacity	0.53	0.65	5.4	12.7k

B. Candidate coverage and newly selected correct answers

Policy	Initial coverage	Final coverage	New-correct	EM
Singleton-first	0.61	0.74	0.06	0.48
Greedy	0.61	0.77	0.07	0.50
Uniform random	0.61	0.70	0.04	0.45
CWE	0.61	0.80	0.09	0.53

C. Initial versus final menu on the same CWE prefix

Selectable menu	EM	F1
C_0	0.50	0.62
C_T	0.53	0.65

B.6 Support structure

Table 6 compares a shared necessary bridge, disjoint alternative supports, and a missing necessary hop, with 150 cases each. Full-context agreement A_V is the modal vote fraction; FC_V is the false-consensus rate. Among EM-correct answers, W_y is the largest supporting-subset count and K_y the largest capacity. The strongest incorrect capacity is $K_{\neg y}$. These scores use exhaustive enumeration and empty-control zeroing; EM measures CWE's online selection.

Disjoint alternatives yield lower initial agreement but higher capacity and EM than the shared bridge. Semantic checks accept 441 of 450 condition instances. Behavioral realization rates are .83, .79, and .91 for the shared bridge, disjoint alternatives, and missing hop, respectively. Because initial agreement also changes, the paired contrasts describe the complete structural intervention.

Table 6: Support structure: condition means (A) and paired contrasts with the shared bridge (B). Brackets give 95% intervals; the capacity margin is $K_y - K_{\neg y}$.

A. Condition means							
Condition	n	A_V	FC_V	W_y	K_y	$K_y - K_{\neg y}$	EM
Shared bridge	150	0.78	0.31	1.2	1.1	+0.2	0.41
Disjoint alternatives	150	0.52	0.12	2.6	2.4	+1.3	0.58
Missing hop	150	0.66	0.44	0.8	0.6	-0.4	0.29

B. Paired changes from the shared-bridge condition			
Contrast	ΔK_y	Δmargin	ΔEM
Alternatives	$+1.30\ [+1.00, +1.60]$	$+1.10\ [+0.80, +1.40]$	$+0.17\ [+0.11, +0.23]$
Missing hop	$-0.50\ [-0.70, -0.30]$	$-0.60\ [-0.80, -0.40]$	$-0.12\ [-0.18, -0.06]$

B.7 When witness exchange matters

The exchange experiment tests whether replacing an early witness helps recover a better packing. Each condition uses 120 questions, three initial-witness variants, and two identifier/order variants. Policies share a paid initial prefix and exact capacity aggregation. Restricted keeps the initial witness and queries only unused evidence. Greedy's fallback can revisit occupied evidence, making it a different comparator.

The allowance is eight, including the common prefix. Table 7 reports the paired interaction: (CWE minus comparator) under exchange-dependent structures minus the same contrast under non-obstructing structures. CWE gains more over Restricted when the initial witness must be replaced.

Table 7: Witness exchange: policy results (A) and paired interactions (B). G_{sel} is the exact-capacity selection gap; brackets give 95% intervals.

A. Policy results by structural condition						
Condition	Policy	EM	F1	Reports	Tokens	G_{sel}
Exchange-dependent	Restricted	0.43	0.55	7.1	13.6k	1.3
Exchange-dependent	Singleton-first	0.45	0.57	6.2	12.9k	1.1
Exchange-dependent	CWE	0.52	0.64	5.6	12.8k	0.5
Non-obstructing	Restricted	0.50	0.62	7.0	13.4k	0.8
Non-obstructing	Singleton-first	0.50	0.62	6.1	12.8k	0.9
Non-obstructing	CWE	0.52	0.64	5.5	12.6k	0.6

B. Exchange-dependent minus non-obstructing policy contrasts		
Comparator	EM interaction [95% CI]	F1 interaction [95% CI]
Restricted	$+0.070\ [+0.020, +0.120]$	$+0.070\ [+0.020, +0.120]$
Singleton-first	$+0.050\ [+0.000, +0.100]$	$+0.050\ [-0.010, +0.110]$

B.8 Certificates and observed errors

Table 8 separates the main panel, exhaustive reference, and reviewed errors. Certified selections have conditional EM of .850: a certificate establishes a capacity winner among discovered candidates, which can still be incorrect. Rescoring the same queries without empty-answer zeroing raises EM by .018 $[.008, .028]$, with the online query sequence fixed. This exposes the cost of excluding useful prior knowledge. Among the 300 reviewed errors, missing-evidence and normalization labels can overlap.

Table 8: Certification and observed errors. Each row states its denominator. The certified-error rate is conditional on certification; the exhaustive-reference and reviewed-error rows use separate subsets.

Event	Count	Denominator	Rate
Main panel: certification and prior exclusion			
Certified selection	4,500	9,000	0.500
Incorrect certified selection	675	4,500	0.150
Empty correct, final CWE wrong	540	9,000	.060
Without prior exclusion: wrong to correct	410	9,000	.046
Without prior exclusion: correct to wrong	250	9,000	.028
Exhaustive-reference subset			
Wrong selection with $K(\hat{a}) > K_y$	117	1,800	.065
Reviewed CWE errors			
Missing necessary evidence in reviewed errors	96	300	.320
Normalization error in reviewed errors	27	300	.090

C COMMUNICATION DIAGNOSTIC BEHIND THE MOTIVATION

The communication diagnostic asks whether agreement can increase without additional evidence. In the earlier five-agent study, communication raises consensus and false-consensus in all 18 model–dataset–seed runs, while EM decreases in 15. False-consensus means that at least four agents agree on a wrong answer.

To isolate the effect of displayed lineage, a matched-text intervention repeats one snippet in ten slots. The slots receive either one shared root ID or ten distinct IDs; their evidence text stays fixed. The case-paired effect compares the communication change between these two conditions: $(R2 - R0)_{\text{distinct-ID}} - (R2 - R0)_{\text{same-root}}$. Across 100 cases, the false-consensus effect is $+.0389$ $[.0156, .0611]$. A separately selected second 50-case cohort gives $+.0489$ $[.0156, .0822]$.

Table 9 shows that the pooled effect is concentrated in Llama; the Qwen and Mistral intervals include zero. The root IDs are constructed lineage labels, so they measure sensitivity to displayed provenance rather than establish independent authorship. The communication protocol and lineage controls appear in the accompanying diagnostic supplement (Sections S1–S2).

Table 9: Effect of distinct lineage labels at fixed evidence text. Each estimate is the difference between the R0–R2 false-consensus changes in the distinct-ID and shared-root conditions. Brackets give paired 95% case-bootstrap intervals.

Backbone	Effect	95% interval
Qwen	$+.013$	$[-.010, .040]$
Llama	$+.110$	$[.053, .167]$
Mistral	$-.007$	$[-.047, .033]$