

Where Do LLM Ensemble Gains Come From? Decomposing Candidate Value, Diversity, and Aggregation

Donghwan Kim

Independent Researcher

Abstract

LLM ensemble gains are often measured against a single baseline, but the value attributed to additional candidates can change with the comparison used. We study an LLM aggregator under four matched conditions: *Direct* (no candidates), *Single-slot* (one candidate in a multi-slot interface), *Duplicate* (one candidate repeated), and *Multi-candidate* (separately generated candidates). Under either non-diverse control, overall candidate gain (CIV) splits algebraically into an additional-candidate contrast (DSV) and a candidate-context contrast (CCV); the split, but not CIV, depends on the chosen control. On MultiHopRAG model diversity ($k=2$), the repetition gap $\text{Rep} = -8.33$ pp $[-13.25, -3.75]$ quantifies the difference between gains measured against Duplicate and Single-slot. On a matched FRAMES stronger-panel extension ($n=200$, $k=3$, aggregator output limit 4096), paired $\Delta\text{Rep} = +9.67$ pp $[+6.00, +13.50]$; within panels, Rep changes sign (original -6.33 ; stronger $+3.33$). In FRAMES evidence aggregation, candidate-context contrasts remain strongly negative under both non-diverse controls. Our results show why ensemble evaluations should state the reference control before attributing gains to additional candidates.

1 Introduction

Work on LLM ensembles usually asks whether more samples, models, or evidence raise final accuracy (Wang et al., 2023; Du et al., 2023; Wang et al., 2024; Li et al., 2024). Beyond majority voting, Universal Self-Consistency uses an LLM to select among free-form candidate answers (Chen et al., 2023), while LLM-Blender ranks and fuses outputs from different models (Jiang et al., 2023). These methods show that combining candidates involves a decision about how their information is used. Adding models is not always beneficial (Li et al., 2025; Kim et al., 2026). Multi-agent debate

also does not consistently outperform simpler self-consistency or ensembling methods (Smit et al., 2024). This makes it important to distinguish the value of additional candidates from the behavior of the system that combines them. A separate audit finds that common diversity metrics often re-express capability, and that latent complementarity rarely becomes majority-vote gain (Kim, 2026a). In tool-mediated industrial diagnosis, specialist information likewise need not turn into a matching gain for an LLM integrator (Kim, 2026c).

Overall accuracy alone does not reveal why an ensemble succeeds or fails. This is a longstanding issue in ensemble learning. Classical work relates ensemble error to diversity among predictors (Krogh and Vedelsby, 1994; Brown et al., 2005). More recently, Choi et al. (2025) separate the effects of voting and debate, while Ali (2026) decomposes majority-vote gains into rescued and damaged predictions. These studies show why aggregate performance can hide important differences between ensemble mechanisms. We do not claim to be the first decomposition of ensemble gains. Our contribution is a matched set of *answer-level* controls for an LLM aggregator that can generate a new answer rather than only selecting a member prediction. We test whether Duplicate and Single-slot, two plausible non-diverse controls, yield the same attribution of candidate value; they do not, and their accuracy gap changes sign between the evaluated candidate-generation configurations.

We organize the measurement around four questions:

1. Does a Multi-candidate panel help relative to *Direct* solving (no candidate answers)?
2. How does Multi-candidate compare with *Single-slot*, which places one candidate in the same multi-slot interface?
3. How does that comparison change when the baseline is *Duplicate*, which repeats one candi-

date across slots?

4. What does the gap between Duplicate and Single-slot reveal about the attributed value of additional candidates?

The central empirical finding is that **the value attributed to additional candidates depends on the non-diverse control**.

Two settings make the dependence concrete. On MultiHopRAG model diversity ($k=2$), gain over Duplicate is large ($DSV_{\text{dup}} = +9.3 \text{ pp}$ [$+5.0, +14.0$]) while overall candidate gain over Direct is -3.0 pp (CI includes zero). Under Single-slot, the same panel yields only $DSV_{\text{slot}} = +1.00 \text{ pp}$ [$-2.33, +4.67$], matching a negative repetition gap ($\text{Rep} = -8.33 \text{ pp}$ [$-13.25, -3.75$]). On FRAMES, a matched stronger-candidate extension ($n=200$) raises total candidate gain by about $+20.5 \text{ pp}$ while changing the estimated gain of additional candidates by $+10.3 \text{ pp}$ under Single-slot but only $+0.7 \text{ pp}$ under Duplicate; within panels, Rep changes sign (original -6.33 ; stronger $+3.33$), with paired $\Delta\text{Rep} = +9.67 \text{ pp}$ [$+6.00, +13.50$]. Separately, evidence-mediated candidate context remains strongly negative under both controls at a 4096-token aggregator ceiling.

Contributions.

1. **Controlled decomposition.** Matched Direct / Single-slot / Duplicate / Multi-candidate conditions (experiment IDs C2 / C1_{slot} / C3b / C1) separate overall candidate gain (CIV), baseline-specific diversity contrasts (DSV_{dup} , DSV_{slot}), candidate-context contrasts (CCV), and the repetition gap (Rep), with elementary identities $\text{CIV} = \text{DSV} + \text{CCV}$ and $DSV_{\text{dup}} = DSV_{\text{slot}} - \text{Rep}$. The accounting is algebraic; the contribution is the empirical controls and their interpretation.
2. **Control dependence with a matched stress test.** On MultiHop model diversity and on a matched FRAMES stronger-panel extension ($n=200$, 4096-token aggregation), the estimated value of additional candidates changes with the non-diverse control. We do not claim Single-slot is uniquely neutral or isolates a pure causal diversity effect.
3. **Effects beyond baseline choice.** Evidence CCV_{slot} and CCV_{dup} remain strongly negative under both controls; raw-evidence restoration and aggregator-configuration comparisons identify losses that are not explained by choosing

Duplicate versus Single-slot.

2 Related Work

Self-consistency, Universal Self-Consistency, multi-agent debate, generative fusion, sampling-and-voting, and mixture-of-agents all combine multiple generations (Wang et al., 2023; Chen et al., 2023; Du et al., 2023; Jiang et al., 2023; Li et al., 2024; Wang et al., 2024). Self-MoA and related work ask whether mixing models helps when proposer quality varies (Li et al., 2025; Abdulaal et al., 2025). Multi-agent debate does not reliably beat simpler ensembling unless carefully tuned (Smit et al., 2024), and voting can account for much of the gain attributed to debate (Choi et al., 2025; Kaesberg et al., 2025).

Classical ensemble theory relates accuracy to disagreement among predictors (Krogh and Vedelsby, 1994; Brown et al., 2005; Kuncheva and Whitaker, 2003); recent LLM work likewise finds that diversity measures need not predict ensemble accuracy (Kim, 2026a; Ali, 2026; Chen, 2026). Kim (2026a) examines diversity metrics as predictors of majority-vote gain, whereas we study how non-diverse aggregation controls affect the attribution of candidate value. We do not claim the same bias-variance-covariance algebra for LLM answer accuracy; we use those results only as motivation for separating mechanism-level contrasts from a single aggregate gain.

Selector quality can matter more than diversity in generate-then-select pipelines (Maryansky et al., 2026). Other work studies trace-level synthesis in mixture-of-agents settings (Fadnavis et al., 2026). LLM-as-a-judge is widely used (Zheng et al., 2023), but producing candidates and selecting among them remain distinct capabilities. Prior work has also documented limitations of judge scores as optimization signals in a separate table-recognition setting (Kim, 2026b). Designed evidence asymmetry can help when deliberation is built for it (Li et al., 2026); long-context work shows that available information need not be used (Liu et al., 2024). In RAG-style aggregation of multiple evidence documents, models can favor repeated paraphrases and presentation order over independent support (Naphade, 2026). That line of work studies how models weight repeated evidence. We instead measure how candidate-level value attribution changes under matched Duplicate and Single-slot *answer* controls. We add a matched

Direct / Single-slot / Duplicate / Multi-candidate contrast that separates candidate-mediated context from baseline-specific diversity increments.

3 Decomposing Ensemble Value

3.1 Four aggregation conditions

Fix one question, a panel of size k , and an aggregator A . Figure 1 shows four matched input constructions (reader-facing names first; experiment IDs in parentheses):

- **Direct** (C2): A solves the question with no candidate answers. In sampling/model settings the prompt is closed-book; in evidence settings the frozen primary Direct condition is a matched evidence-conditioned direct solve.
- **Single-slot** (C1_{slot}): the same k -slot candidate interface with one filled candidate and $k-1$ empty slots. For $k=3$, the primary construction is $[c_i, \text{empty}, \text{empty}]$ with the first slot filled and placeholder “(No answer was provided for this slot.)”. Reported accuracy averages over candidate index i , not over slot position (position and placeholder wording are later sensitivities).
- **Duplicate** (C3b): k constructed copies of one candidate, $[c_i, c_i, c_i]$ for $k=3$, averaged over candidate index i .
- **Multi-candidate** (C1): k separately generated candidate blocks. Final answer strings may still coincide.

Aggregator prompts for Single-slot, Duplicate, and Multi-candidate are candidate-only: raw evidence is not reinjected into those arms. Sampling, evidence, and the matched stronger-candidate extension use $k=3$. The MultiHop and FRAMES model-diversity results we highlight use a pairwise ($k=2$) unit.

3.2 Overall candidate gain

CIV (candidate-information value) is Multi-candidate minus Direct:

$$\text{CIV} = \text{Acc}(\text{C1}) - \text{Acc}(\text{C2}). \quad (1)$$

It is the accuracy difference relative to the matched Direct condition. The difference can be positive, near zero, or negative; negative CIV means the Multi-candidate setup underperforms Direct on that unit.

3.3 Gains over alternative non-diverse controls

DSV_{slot} is Multi-candidate minus Single-slot:

$$\text{DSV}_{\text{slot}} = \text{Acc}(\text{C1}) - \text{Acc}(\text{C1}_{\text{slot}}). \quad (2)$$

It asks how much accuracy changes when separately generated candidates replace one filled slot in the same interface. Adding candidates also changes how much candidate text is present, so this is not a pure causal diversity estimand.

DSV_{dup} is Multi-candidate minus Duplicate:

$$\text{DSV}_{\text{dup}} = \text{Acc}(\text{C1}) - \text{Acc}(\text{C3b}). \quad (3)$$

It asks how much accuracy changes when separately generated candidates replace constructed repetitions of one answer. The two baselines answer different questions; both are useful, and neither is uniquely “the” diversity effect. In historical primary tables, bare “DSV” means DSV_{dup} .

3.4 The repetition gap

Rep is Duplicate minus Single-slot:

$$\text{Rep} = \text{Acc}(\text{C3b}) - \text{Acc}(\text{C1}_{\text{slot}}). \quad (4)$$

In plain language, the repetition gap measures how much accuracy changes when one candidate is repeated instead of shown once in the multi-slot interface. Negative Rep means Duplicate underperforms Single-slot; positive Rep means Duplicate outperforms Single-slot. The baseline dependence of the two diversity contrasts is explicit in the identity

$$\text{DSV}_{\text{slot}} - \text{DSV}_{\text{dup}} = \text{Rep}, \quad (5)$$

equivalently $\text{DSV}_{\text{dup}} = \text{DSV}_{\text{slot}} - \text{Rep}$.

3.5 Candidate-context values

After the comparisons above, we define candidate-context contrasts versus Direct:

$$\text{CCV}_{\text{slot}} = \text{Acc}(\text{C1}_{\text{slot}}) - \text{Acc}(\text{C2}), \quad (6)$$

$$\text{CCV}_{\text{dup}} = \text{Acc}(\text{C3b}) - \text{Acc}(\text{C2}). \quad (7)$$

Accounting identities then split the same CIV under either baseline:

$$\text{CIV} = \text{DSV}_{\text{dup}} + \text{CCV}_{\text{dup}} = \text{DSV}_{\text{slot}} + \text{CCV}_{\text{slot}}. \quad (8)$$

CIV itself does not depend on choosing Duplicate versus Single-slot; what changes is the DSV/CCV split. We call these terms **candidate-context value** because they mix candidate quality, text presence, repetition or interface structure, and aggregator reaction. They are accounting contrasts, not pure causal mediators.

Four ways to aggregate candidate answers ($k=3$; comparisons, not execution order)

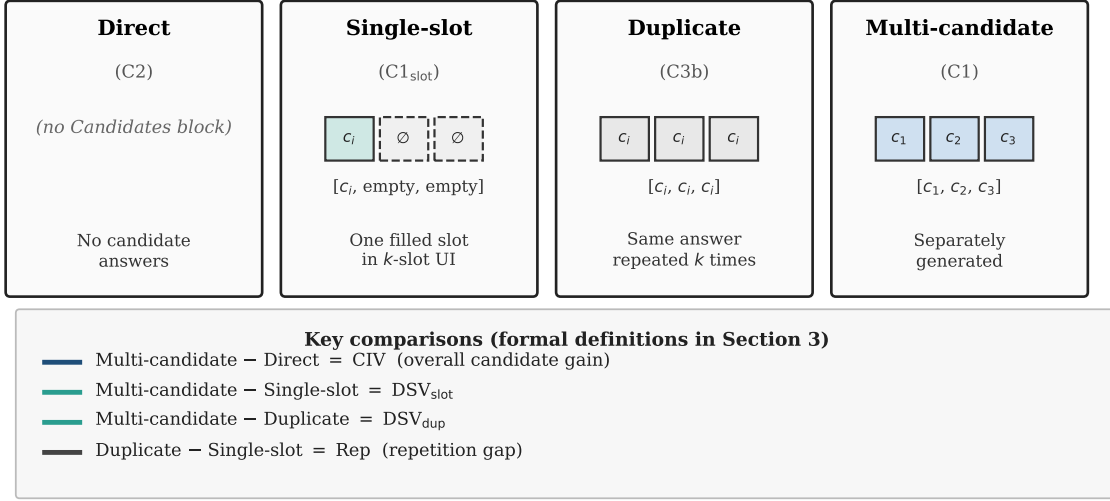


Figure 1: Four aggregation conditions for panel size $k=3$ (comparisons, not an execution order). Reader-facing names are primary; experiment IDs appear in parentheses. Single-slot and Duplicate accuracies are averaged over candidate index i ; Single-slot does not average over slot position in the primary Acc. Separately generated Multi-candidate answers may coincide. Neither non-diverse control identifies pure causal diversity. Formal estimands are defined in Section 3.

3.6 What the measures do not identify

Repetition changes the input interface and can influence aggregator behavior. Single-slot changes candidate amount and coverage relative to Multi-candidate. Separately generated candidates can differ in quality, correctness, and content. Neither DSV contrast alone identifies pure causal diversity.

4 Experimental Setup

Benchmarks and units. FRAMES (Krishna et al., 2025) (sampling $n=200$; evidence $n=180$) and MultiHopRAG (Tang and Yang, 2024) (replication $n=100$). FRAMES model diversity uses a pairwise analysis covering 30 distinct items, not the generic FRAMES $n=200$ pool. MultiHop model diversity averages three fixed $k=2$ model-pair comparisons within each of $n=100$ items.

Sources and aggregator. Sampling (S), model (M), and evidence (E) diversify candidates. Sampling diversity uses three samples per generator, averaged across three generators. For evidence diversity on both benchmarks, candidates come from a fixed Qwen3.8-Flash generator, and the Direct reference is the matched evidence-conditioned Direct, rather than union-evidence or closed-book Direct. The main aggregator (“Luna”) sees parsed final answers. Many confirmatory runs use a 1024-token output limit; later checks and the matched stronger-

candidate extension use 4096. A Qwen3-32B extension keeps candidate content fixed and swaps the aggregator *configuration* (model plus decoding and reasoning settings), not a model-only intervention with equal decoding. Verified model labels, decoding settings, and prompt locations appear in Appendix M. Table 4 (Appendix B) summarizes datasets, panel sizes, token ceilings, and analysis status.

Scoring and analysis status. We use exact/normalized string match, then a frozen semantic judge (Zheng et al., 2023). Unless noted, we report paired complete-case estimates with bootstrap 95% CIs ($B=2000$, seed 2026 where specified); no p -values. Historical primary analyses, later controlled extensions, and exploratory diagnostics are distinguished in Appendix C (Table 5); model settings and prompt sources appear in Appendix M. The historical primary S/M/E analyses use $DSV = DSV_{dup}$. We do not relabel all later results as confirmatory.

Human audit. We audited a late-stage stratified sample of $n=60$ judged responses (seed 2026; twelve cases from each of five strata), not a simple random sample of the completed judged universe. Two independent raters adjudicated eight disagreements. Versus final human labels, the judge agreed on 50/60 (83.3%; $\kappa=0.645$). False rejection

tions on this packet are mostly paraphrase or alias mismatches on FRAMES aggregator answers. The audit does not support a CIV/DSV/CCV correction (Appendix E; Tables 7 and 8).

5 Control-Dependent Ensemble Value

Table 1 reports the historical Duplicate-relative decomposition (plotted in Appendix A), motivating the matched Single-slot and repetition-gap analyses that follow. Evidence CIV (Multi-candidate minus Direct) is strongly negative on both benchmarks, while evidence DSV_{dup} is near zero: most of the loss is candidate-context value, not a baseline-specific diversity increment. This differs from settings where designed evidence asymmetry is intended to help deliberation (Li et al., 2026). On MultiHop models, $DSV_{dup} = +9.3$ pp while CIV is -3.0 pp: a positive gain over Duplicate does not imply a positive candidate gain over Direct. Sampling effects are smaller and task-dependent (near zero CIV on FRAMES; negative CIV on MultiHop). We therefore first test whether the Duplicate baseline itself can depress accuracy, then test whether the pattern changes under a matched stronger-candidate extension.

Control dependence. Under the same MultiHop model panel ($n=100$), gain over Single-slot is only $DSV_{slot} = +1.00$ pp $[-2.33, +4.67]$, while the repetition gap is $Rep = -8.33$ pp $[-13.25, -3.75]$, so $DSV_{dup} = DSV_{slot} - Rep$ recovers $+9.33$ pp (Figure 2a). Under negative Rep, DSV_{dup} exceeds DSV_{slot} by $-Rep$. This does not make DSV_{dup} wrong; it measures a different contrast. A CI that includes zero means the contrast is unresolved at this precision, not that the effect is zero. The MultiHop model-diversity result is control-sensitive: DSV_{dup} and DSV_{slot} tell different stories. We next ask whether the same control dependence appears when the candidate-generation configuration is strengthened on FRAMES.

Stronger candidates under two controls. Is a small Duplicate-relative DSV only a symptom of weak candidates? A weak panel can make both total CIV and diversity-related increments look small for uninteresting reasons. We therefore ask whether raising candidate quality changes the picture when the aggregator and item set stay matched. We compare an original candidate panel with a higher-performing panel generated using different models and generation settings. We hold the questions,

panel size ($k=3$), aggregator (Luna), and aggregator output limit (4096 tokens) fixed, on $n=200$ FRAMES items (the original 95-item cohort plus a completed nonoverlapping 105-item complement). The original panel uses Solar Mini4, Qwen3.8-Flash, and Haiku 4.5 with temperature 0.0 and a 1024-token candidate-output limit (no explicit reasoning effort set). The stronger panel uses Gemini 3.8 Flash, Grok 4.7, and Sonnet 5.5 with temperature 0.0, reasoning effort high, and a 65536-token candidate-output limit. The 4096-token aggregation limit is distinct from these candidate-generation limits. The resulting contrast measures the value of replacing the candidate-generation configuration, not the isolated effect of model identity under equal candidate-generation compute.

Contrasts are stronger minus original (paired bootstrap, $B=2000$, seed 2026). Improving the candidate-generation configuration raises CIV by $+20.50$ pp $[+14.00, +27.00]$. Table 2 reports the absolute arm accuracies, which clarify the within-panel signs. Table 3 reports the matched changes. On the matched FRAMES items, the original Multi-candidate panel is below Direct (41.5% versus 46.5%; $CIV \approx -5.0$ pp), whereas the stronger panel is above Direct (62.0% versus 46.5%; $CIV \approx +15.5$ pp). This reversal corresponds to the paired $\Delta CIV = +20.50$ pp, while the attributed additional-candidate contrast changes by $+0.67$ pp under Duplicate versus $+10.33$ pp under Single-slot. The key paired contrast is $\Delta Rep = +9.67$ pp $[+6.00, +13.50]$: it directly shows that the two diversity attributions differ, rather than relying on one ΔDSV CI excluding zero and another including zero (Figure 2b; Figure 3). On the original panel, $Rep = -6.33$ pp $[-9.33, -3.67]$ (Duplicate worse than Single-slot). On the stronger panel, $Rep = +3.33$ pp $[+0.83, +5.83]$ (Duplicate better than Single-slot). Thus $\Delta DSV_{slot} = +10.33$ pp and $\Delta DSV_{dup} = +0.67$ pp are different splits of the same ΔCIV , linked by ΔRep . Context shifts are $\Delta CCV_{dup} = +19.83$ pp $[+15.17, +24.67]$ and $\Delta CCV_{slot} = +10.17$ pp $[+6.33, +14.17]$.

Answer-count sensitivity. A separate controlled intervention on eligible mixed-answer cases finds that changing answer multiplicity changes aggregation accuracy by $+8.12$ pp $[+4.99, +11.42]$ ($n=197$). This supports sensitivity to answer counts but does not establish the mechanism behind the repetition-gap differences observed across candidate-generation configurations.

Table 1: Main Luna breakdown (pp, 95% CIs). DSV is gain over Duplicate (DSV_{dup} ; C3b control). FRAMES M uses a pairwise $k=2$ unit (30 distinct items), not FRAMES $n=200$. Evidence CCV uses mean-duplicate C3b.

Bench.	Src.	CIV	DSV_{dup}	CCV_E
FRAMES	S	-0.8 [-5.2, 4.0]	0.0 [-2.7, 3.0]	n/a
FRAMES	M	+0.7 [-10.3, 12.0]	+1.8 [-2.5, 6.8]	n/a
FRAMES	E	-15.9 [-22.6, -9.4]	+1.9 [-0.9, 4.6]	-17.8 [-24.1, -11.3]
MultiHop	S	-6.0 [-11.3, -0.7]	+4.3 [0.7, 8.0]	n/a
MultiHop	M	-3.0 [-9.0, 3.0]	+9.3 [5.0, 14.0]	n/a
MultiHop	E	-13.3 [-20.7, -6.3]	+0.1 [-2.0, 2.1]	-13.4 [-20.3, -6.6]

The $n=95/n=105$ rows are same-protocol diagnostics, not independent benchmarks (Appendix D, Table 6). The MultiHop model-diversity contrast used pairwise $k=2$; the matched FRAMES extension uses $k=3$, so they are not independent replications of one intervention. Earlier standalone-candidate evaluations used a different stack and are reported separately in Appendix H; their estimates are not pooled with the matched comparison. An exploratory candidate-conditioned breakdown suggests that repetition effects vary with candidate correctness; all four correctness-by-panel cells are reported in Appendix L. Correctness is post-generation and is not a causal mediator. Slot-prompt position and placeholder variants on an $n=60$ subset keep positive ΔDSV_{slot} point estimates while variant-minus-primary CIs include zero (Appendix K). The Single-slot contrast measures incremental value over a single-filled-slot control, not pure causal diversity.

Table 2: Absolute arm accuracies (%) and Rep (pp) on matched FRAMES $n=200$ (aggregator output limit 4096). Direct (C2) is shared across panels. Columns: Direct / Single-slot / Duplicate / Multi-candidate (experiment IDs C2 / C1_{slot} / C3b / C1). Rep = Duplicate - Single-slot. Absolute accuracies are rounded to one decimal; Rep uses the unrounded matched contrast. Rep CIs are item-paired bootstrap ($B=2000$, seed 2026).

Panel	Direct	Single-slot	Duplicate	Multi-cand.	Rep
Original	46.5	45.2	38.8	41.5	-6.33 [-9.33, -3.67]
Stronger	46.5	55.3	58.7	62.0	+3.33 [+0.83, +5.83]

6 Information Loss Beyond Baseline Choice

Lost information through candidates. Making information available is not the same as extracting it into candidate answers, and extracting it is not the same as the aggregator using it (Liu et al., 2024). On FRAMES, candidates-only accuracy (47.2%) is far below RAW (65.6%), and RAW+candidates

(65.7%) returns almost to RAW. On MultiHop, recovery is only +6.0 pp (91.0 / 74.7 / 80.7), so candidate context can still hurt (Figure 4a). We do not claim a position-bias mechanism.

Output-limit check. Re-running FRAMES evidence aggregation at 4096 tokens ($n=180$) lowers Multi-candidate evidence truncation from 20.4% to 1.7%, but CIV_E moves from -15.9 pp to -18.33 pp [-24.26, -12.03]. $CCV_{slot} \approx -17.56$ pp and $CCV_{dup} \approx -18.70$ pp remain strongly negative under both baselines (Figure 4b). CIV itself is unchanged by the choice of non-diverse baseline; the empirical finding is that both candidate-context contrasts stay strongly negative. Short outputs do not explain the evidence gap. On robustness subsets, a second aggregator family shows the same qualitative pattern: evidence CIV stays strongly negative with near-zero evidence DSV, while model DSV can be positive (Appendix G, Table 9).

Control robustness. The remaining evidence checks distinguish information preserved in raw evidence, extracted into candidates, and used by the aggregator. The corresponding evidence repetition gap is small at the observed point estimate, $Rep \approx -1.15$ pp [-4.23, +2.01] (Appendix J). This interval does not establish that Duplicate and Single-slot are equivalent.

Additional checks. Visible full explanations do not clearly raise DSV_{dup} in our checks (FRAMES -0.67; MultiHop -2.33; both CIs include zero); a compact-rationale follow-up is similarly unresolved (Appendix I, Table 11). Trace-level synthesis is methodologically different from our answer-level checks (Fadnavis et al., 2026). These secondary checks are reported in the appendix and do not change the evidence-loss result.

Table 3: Matched FRAMES stronger-minus-original effects (pp; $k=3$, aggregator output limit 4096). Pooled $n=200$ is primary. $\Delta\text{CCV}_{\text{dup}} = +19.83$ [$+15.17, +24.67$] and $\Delta\text{CCV}_{\text{slot}} = +10.17$ [$+6.33, +14.17$] are stated in text and Figure 3. Cohort diagnostics are in Appendix D.

	ΔCIV	$\Delta\text{DSV}_{\text{dup}}$	$\Delta\text{DSV}_{\text{slot}}$	ΔRep
Pooled $n=200$	+20.50 [14.00, 27.00]	+0.67 [−6.00, 7.00]	+10.33 [3.67, 16.67]	+9.67 [6.00, 13.50]

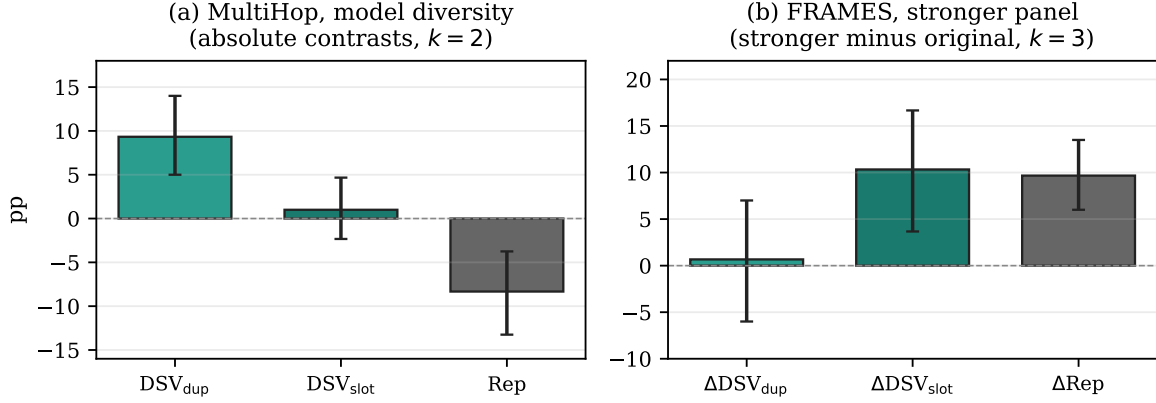


Figure 2: Left: absolute MultiHop model-diversity contrasts (gain over Duplicate, gain over Single-slot, and Rep). Right: matched FRAMES stronger-minus-original changes ($n=200$, 4096-token aggregation). Both panels show paired 95% CIs.

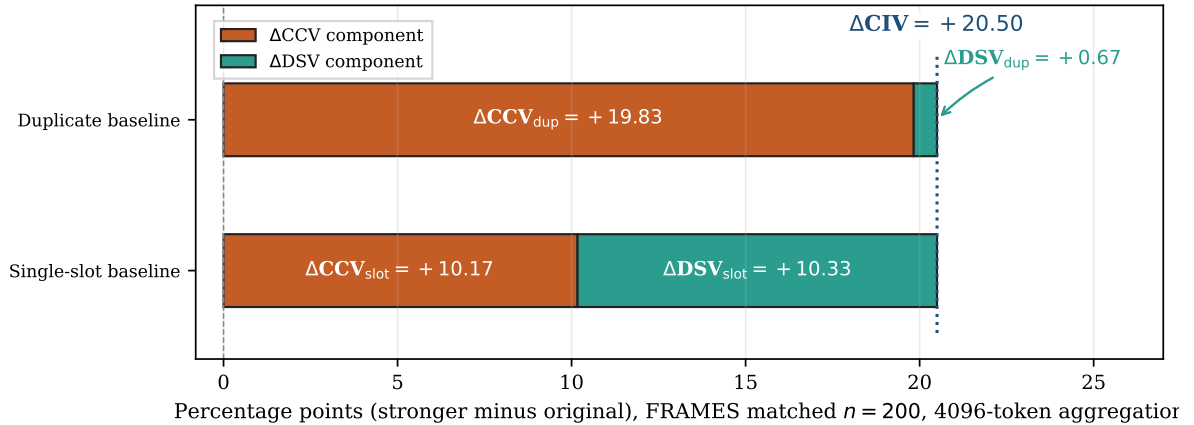


Figure 3: Accounting split of the same ΔCIV under Duplicate and Single-slot (not a causal path diagram). Duplicate: $\Delta\text{CCV}_{\text{dup}} = +19.83$, $\Delta\text{DSV}_{\text{dup}} = +0.67$; Single-slot: $\Delta\text{CCV}_{\text{slot}} = +10.17$, $\Delta\text{DSV}_{\text{slot}} = +10.33$. Both sum to $\Delta\text{CIV} = +20.50$; $\Delta\text{Rep} = +9.67$ links the DSV values. Paired CIs are in Table 3.

7 Discussion and Conclusion

The results imply several practices for ensemble evaluation and design.

Always report the baseline. A diversity-related gain should state whether it is measured over Single-slot or Duplicate. On MultiHop model diversity ($k=2$) and the matched FRAMES stronger-panel extension ($k=3$), DSV_{dup} and DSV_{slot} tell different design stories. If gain over Duplicate is large while gain over Single-slot is not, adding

more models is not automatically justified by the Duplicate-relative number alone.

Do not assume repetition is neutral. Matched FRAMES results show that Rep can change sign between candidate-generation configurations. This does not establish a universal causal mechanism for the sign change.

Candidate quality and diversity attribution are separate questions. Improving the candidate-generation configuration can raise CIV without

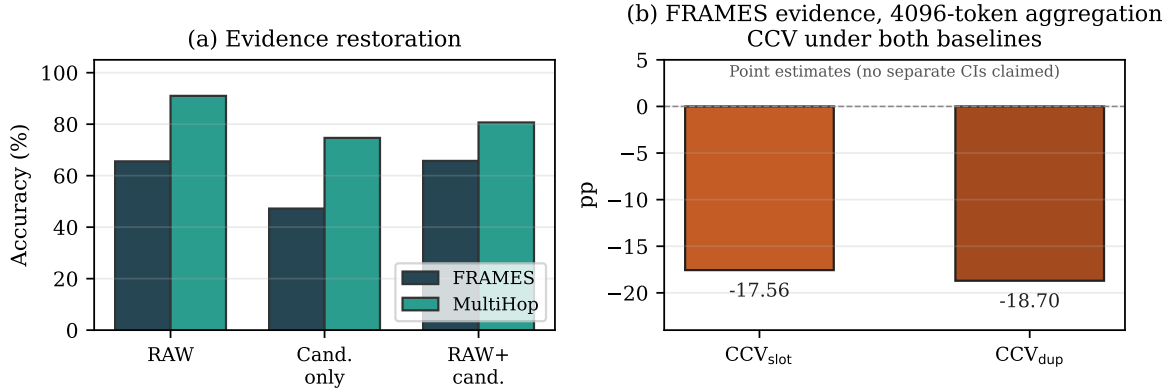


Figure 4: Left: restoring raw evidence beside candidates (accuracy %). Right: FRAMES evidence candidate-context contrasts under Single-slot and Duplicate at 4096 tokens (point estimates). CIV (Multi-candidate – Direct) does not change between those two baselines by definition.

producing the same increase under every diversity-related contrast.

Restoring raw evidence recovers most of the FRAMES gap. Evidence can be lost through candidate answers or aggregation; on FRAMES, restoring RAW beside candidates recovers most of the observed gap. MultiHop recovery is only partial.

Aggregator configuration is a distinct design variable. Holding candidate content fixed, the Luna-versus-Qwen3-32B evidence CIV gap is +12.59 pp [+4.07, +21.11] on FRAMES and +11.0 pp [+4.67, +17.67] on MultiHop. These observed configuration differences show that generating candidates and realizing their information with an aggregator are distinct design questions. They are not equal-decoding/equal-reasoning model-only interventions, do not establish Qwen Rep, and do not imply a universal ordering.

Cost-accuracy implications. Figure 6 (Appendix F) compares the plotted deployment cost and accuracy of the observed policies. In its evidence-backed panel, Direct has higher plotted accuracy at lower plotted cost than the displayed M, E, and M+E candidate-based policies, illustrating why a diversity-related DSV contrast alone does not establish that extra candidate-generation and aggregation calls are worthwhile compared with Direct. These are descriptive within-panel comparisons: policies use different token limits, and C2_{PRO} also changes model setup, so Figure 6 does not establish a general cost-optimal or equal-reasoning-budget policy.

Conclusion. For LLM ensembles, evaluation should state which counterfactual defines the measured gain, how much candidate information survives the interface, and which effects persist across controls. Reporting CIV, baseline-specific DSV/CCV, and Rep makes those distinctions visible.

8 Limitations

We use selected short-answer benchmarks and mainly answer-level aggregation. Visible-explanation checks do not cover hidden traces. Single-slot (C1_{slot}) changes coverage as well as distinctness; Duplicate (C3b) changes repetition; neither is a pure causal diversity estimand. The matched stronger-versus-original contrast changes candidate models and generation settings together, so it is not a model-only effect at equal candidate-generation compute. Evidence-diversity findings use a single fixed candidate generator on both benchmarks, so their transfer to other generators is untested.

Headline tests primarily use one aggregator (Luna); Qwen CIV gaps do not establish Qwen Rep. The human audit is a stratified $n=60$ packet and does not demonstrate absence of arm-differential judge error on the matched Single-slot/Duplicate cells; we do not numerically correct main effects. Aggregation truncation was tracked, and the frozen inclusion policy was retained rather than switching to success-only accuracy. We do not claim a complete candidate-generation empty/parse-failure comparison between panels beyond the disclosed generation-config differences and aggregation-arm health checks.

Candidate-conditioned Rep analyses are exploratory and condition on a post-generation variable. Slot position/placeholder variants ($n=60$) do not prove equivalence of prompt forms. Later extensions were motivated by earlier findings; they are not all confirmatory. Public OpenRouter slugs and decoding settings appear in Appendix M; vendor-internal build IDs remain unresolved.

References

- Ahmed Abdulaal, Chen Jin, Nina Montaña-Brown, Aryo Pradipta Gema, Daniel C. Castro, Daniel C. Alexander, Philip Alexander Teare, Tom Diethe, Dino Oglic, and Amrutha Saseendran. 2025. [Balancing act: Diversity and consistency in large language model ensembles](#). In *International Conference on Learning Representations (ICLR)*.
- Junade Ali. 2026. Quantifying diversity of thought: A predictive law of weighted LLM ensemble lift. *arXiv preprint arXiv:2607.17384*.
- Gavin Brown, Jeremy L. Wyatt, and Peter Tiño. 2005. [Managing diversity in regression ensembles](#). *Journal of Machine Learning Research*, 6:1621–1650.
- Josef Chen. 2026. When does combining language models help? a co-failure ceiling on routing, voting, and mixture-of-agents across 67 frontier models. *arXiv preprint arXiv:2606.27288*.
- Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2023. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*.
- Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. 2025. [Debate or vote: Which yields better decisions in multi-agent large language models?](#) In *Advances in Neural Information Processing Systems*.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.
- Shreyas Fadnavis, Praiatayini Kanakaraj, and Felix Wyss. 2026. Beyond consensus: Trace-level synthesis in mixture of agents. *arXiv preprint arXiv:2605.29116*.
- Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. [Voting or consensus? decision-making in multi-agent debate](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11640–11671, Vienna, Austria. Association for Computational Linguistics.
- Donghwan Kim. 2026a. Are diversity metrics measuring diversity? a capability-controlled audit of majority-vote gain in LLM ensembles. *arXiv preprint arXiv:2607.20768*.
- Donghwan Kim. 2026b. LLM-as-a-judge scores are unreliable optimization signals in closed-loop table recognition. *arXiv preprint arXiv:2607.13347*.
- Donghwan Kim. 2026c. When LLMs sit above diagnostic tools: Unrealized complementarity in industrial fault diagnosis. *arXiv preprint arXiv:2610.05031*.
- Yubin Kim, Ken Gu, Chanwoo Park, Chunjong Park, Samuel Schmidgall, A. Ali Heydari, Yao Yan, Zhihan Zhang, Yuchen Zhuang, Yun Liu, Mark Malhotra, Paul Pu Liang, Hae Won Park, Yuzhe Yang, Xuhai Xu, Yilun Du, Shwetak Patel, Tim Althoff, Daniel McDuff, and Xin Liu. 2026. [Capable language models can outgrow the benefits of collaboration](#). *Nature Machine Intelligence*, 8:1157–1172.
- Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2025. [Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4745–4759, Albuquerque, New Mexico. Association for Computational Linguistics.
- Anders Krogh and Jesper Vedelsby. 1994. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, volume 7, pages 231–238. MIT Press. NIPS 1994.
- Ludmila I. Kuncheva and Christopher J. Whitaker. 2003. [Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy](#). *Machine Learning*, 51(2):181–207.
- Junyou Li, Qin Zhang, Yangbin Yu, Qiang Fu, and Deheng Ye. 2024. [More agents is all you need](#). *Transactions on Machine Learning Research*.
- Wenzhe Li, Yong Lin, Mengzhou Xia, and Chi Jin. 2025. Rethinking mixture-of-agents: Is mixing different large language models beneficial? *arXiv preprint arXiv:2502.00674*.
- Yuante Li, Yicheng Tao, Kate Zhang, Taozhi Wang, Gefei Gu, and Yaxin Zhou. 2026. Diverse evidence, better forecasts: Multi-agent deliberation under information asymmetry. *arXiv preprint arXiv:2607.01661*.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.

- Artem Maryanskyy, Dmitry Budnikov, and Alibek T. Kaliyev. 2026. [When agents disagree: The selection bottleneck in multi-agent LLM pipelines](#). *Applied Sciences*, 16(10).
- Atharv Naphade. 2026. [Rational synthesizers or heuristic followers? analyzing LLMs in RAG-based question-answering](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 40293–40311, San Diego, California, United States. Association for Computational Linguistics.
- Andries Petrus Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D Barrett, and Arnu Pretorius. 2024. [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 of *Proceedings of Machine Learning Research*, pages 45883–45905. PMLR.
- Yixuan Tang and Yi Yang. 2024. [MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries](#). In *Conference on Language Modeling (COLM)*.
- Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Historical Primary Split Figure

Figure 5 visualizes the historical primary S/M/E split summarized in Table 1. It provides a cross-source view of CIV, DSV_{dup}, and evidence CCV; Table 1 remains the source of the reported estimates.

B Experiment Design Summary

Dataset	Source	n	k	Tokens	Unit	Status
FRAMES S		200	3	1024	item	Primary (A)
FRAMES M		30	2	1024	pairwise	Primary (A)
FRAMES E		180	3	1024	item	Primary (A)
MultiHop S/M/E		100	2/3	1024	item/pair	Primary (A)
FRAMES stronger		200	3	4096	item	Extension (B)

Table 4: Core experiment design (verified against frozen configs). FRAMES M uses 30 distinct items with pairwise comparison/averaging, not the generic FRAMES $n=200$ pool. The matched stronger-candidate row combines the original 95-item cohort and a nonoverlapping 105-item complement; it is a later controlled extension, not two independent benchmarks.

For FRAMES evidence, the frozen 180-item panel assigns three per-item budget-matched Wikipedia evidence packets using source-disjoint allocation where feasible; each candidate is generated from one packet, whereas matched Direct receives their deterministic interleaving.

C Analysis-Status Transparency

Analysis	Status	Notes
S/M/E: CIV, DSV _{dup} , CCV	Confirmatory (A)	Historical primary
4096 checks; RAW+candidates; stronger-panel extension; MultiHop Rep/slot; controlled answer-multiplicity; Qwen; explanations	Extension (B)	Fixed before outcomes
Slot $n=60$; compact rationale	Diagnostic (C)	Not primary
Human audit $n=60$	Completed (B)	Stratified; no recalibration

Table 5: Analysis hierarchy.

D Matched Cohort Diagnostics

Cohort	ΔCIV	$\Delta\text{DSV}_{\text{dup}}$	$\Delta\text{DSV}_{\text{slot}}$	ΔRep
Original cohort $n=95$	+17.89 [9.47, 26.32]	-1.05 [-10.18, 8.42]	+10.53 [1.75, 20.00]	+11.58 [6.67, 16.84]
Compl. $n=105$	+22.86 [13.33, 31.43]	+2.22 [-6.67, 10.48]	+10.16 [0.95, 18.73]	+7.94 [2.86, 13.33]

Table 6: Matched FRAMES stronger-minus-original cohort diagnostics (pp). Same protocol cohorts, not independent benchmarks.

The complement completion recorded 13 truncations; scoring retains the matched denominator.

E Human Audit Details

Late-stage stratified sample: twelve cases from each of five judged strata (standard S/M/E, factorial evidence-layout manipulation, MultiHop replication, Kimi robustness, and alternate Direct configuration C2-Pro; seed 2026), drawn from the then-available judged universe and not redrawn after later call completion. Two raters labeled Correct / Incorrect / Ambiguous, blinded to judge decisions; eight disagreements were adjudicated.

Quantity	Value
Rater agreement (3-class)	52/60=86.7%; $\kappa=0.746$
Human vs. judge agreement	50/60=83.3%; $\kappa=0.645$
TP / FP / TN / FN	16 / 0 / 34 / 10

Table 7: Human audit summary (stratified $n=60$).

	Human Correct	Human Incorrect
Judge Correct	TP=16	FP=0
Judge Incorrect	FN=10	TN=34

Table 8: Confusion matrix (positive = judge Correct; reference = final human).

All ten FN cases on this packet are FRAMES aggregator paraphrase, alias, or unit-equivalent mismatches. One FRAMES audit case (A05) was flagged for external AI assistance during labeling; excluding it gives human-judge agreement 49/59 (vs. 50/60 overall), with specificity unchanged. We do not invent additional assistance flags or recalibrate experimental contrasts.

F Cost Summary

Figure 6 presents the cost-accuracy view for the evaluated policies. It summarizes the historical primary policy evaluations, not the matched stronger-candidate extension. It is descriptive because policies use different token limits and C2_{PRO} also changes the model setup.

G Kimi Robustness (Selective)

Bench.	Estimand	Est. (pp)	95% CI
FRAMES	CIV _E	-15.7	[-28.7, -2.8]
FRAMES	DSV _E	+1.5	[-3.7, +7.4]
FRAMES	DSV _M	+6.7	[0.0, +13.3]
MultiHop	CIV _E	-15.0	[-27.2, -3.9]
MultiHop	DSV _E	+0.2	[-2.4, +2.6]
MultiHop	DSV _M	+5.6	[+0.6, +10.6]

Table 9: Kimi robustness subsets (not the main Luna confirmatory panel).

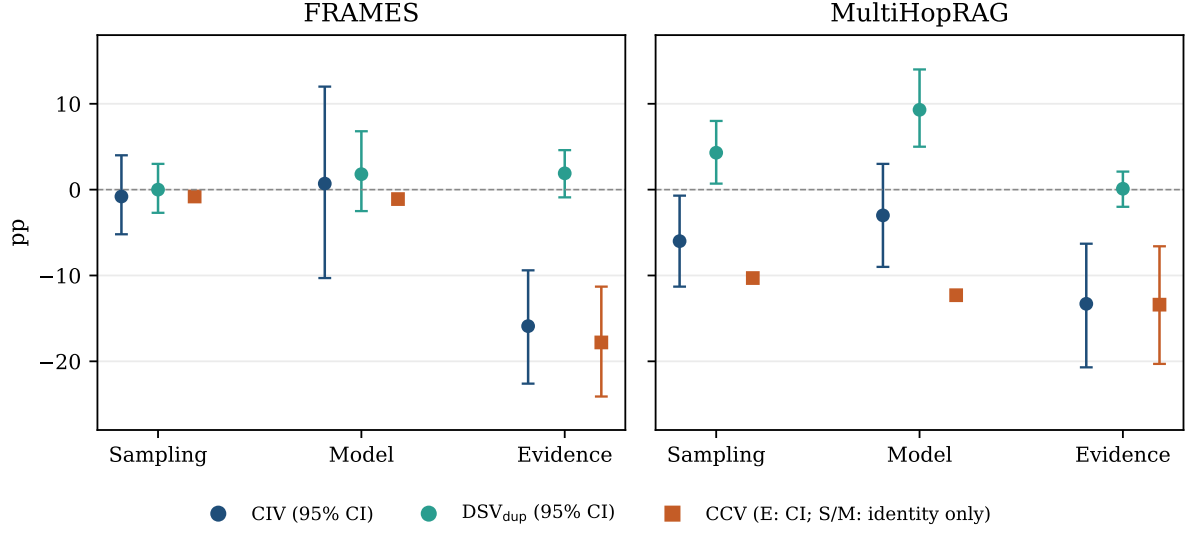


Figure 5: **Appendix A. Historical primary split figure.** DSV = DSV_{dup}. CIV and DSV_{dup} show paired 95% CIs. Evidence CCV has a CI; sampling/model CCV markers are identity points without separate CIs. Moved from the main text because Table 1 already carries the primary numbers.

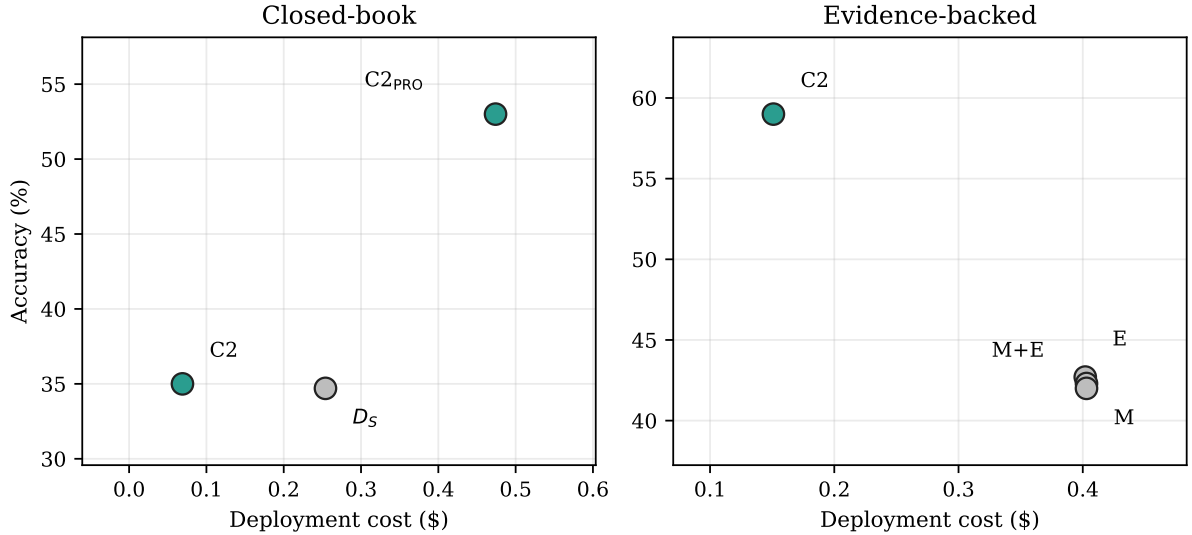


Figure 6: **Appendix F. Cost summary.** Accuracy versus deployment dollar cost. Costs are total deployment costs (USD) on a 100-item FRAMES overlap subset of the historical panels; judge costs are excluded. Because policy construction and model averaging differ from Table 1, these accuracies are not directly comparable to Table 1 estimates. Here D_S denotes sampling diversity: three samples per generator, averaged across the three generators. Filled markers are non-dominated within each panel. Policies use different token limits; C2_{PRO} also changes model setup.

H Task-wise Stronger-Candidate Accuracies

Bench.	Candidate configuration	Acc.	vs. Luna Direct (C2)
FRAMES	Luna Direct (C2)	44.5%	n/a
FRAMES	Gemini High	61.5%	higher
FRAMES	Grok High	52.0%	higher
FRAMES	Sonnet High	50.0%	higher
MultiHop	Luna Direct (C2)	87.0%	n/a
MultiHop	Gemini High	95.0%	higher
MultiHop	Grok High	85.0%	lower
MultiHop	Sonnet High	72.0%	lower

Table 10: Standalone task accuracy vs. Luna Direct (C2) on the earlier standalone-candidate evaluation stack (FRAMES Luna Direct = 44.5%). This is not the matched stronger-panel Direct (C2) in Table 2 (46.5% at aggregator output limit 4096).

I Compact Rationale (Exploratory)

Answer-only (A0) places only the parsed answer in each candidate slot. Compact visible rationale (R1) appends a sentence-boundary-respecting prefix of the visible explanation, truncated to $K=200$ tokens with o200k_base; Full visible explanation (R2) appends the complete visible explanation. This exploratory comparison reuses stored stronger-panel candidate responses and their visible explanations, without new candidate generation or model-based summarization. The original candidate panel fell below the frozen 95% visible-explanation coverage requirement (FRAMES 29%; MultiHopRAG 47%) and was excluded. Here DSV is Duplicate-relative, $DSV_{\text{dup}} = \text{Acc}(C1) - \text{mean}_i \text{Acc}(C3b_i)$, so the rows are changes in that contrast between representations; K is rationale-prefix length, not panel size k .

Contrast	FRAMES	MultiHop
R1-A0 DSV	-1.50 [-8.17, +5.00]	+2.33 [-4.00, +8.67]
R2-R1 DSV	+0.83 [-5.17, +6.67]	-4.67 [-10.67, +0.67]

Table 11: Post-hoc exploratory compact-rationale contrasts ($K=200$). Defined after the full-explanation results; does not change primary verdicts.

J Repetition Gaps Across Settings

Table 2 reports within-panel accuracies on the matched FRAMES stack. Across settings with matched C3b and $C1_{\text{slot}}$, Rep varies in sign and magnitude: MultiHop model diversity Rep = -8.33 pp [-13.25, -3.75] ($k=2$); FRAMES evidence Rep $\approx$ -1.15 pp [-4.23, +2.01] at aggregator output limit 4096 (CI includes zero). These

rows differ in task, k , and panel, so they illustrate sensitivity rather than a universal law.

K Slot-Prompt Sensitivity

On a frozen $n=60$ subset of the matched FRAMES stack, primary $\Delta DSV_{\text{slot}} = +7.50$ pp [-2.22, +17.50]. Filling the last slot instead yields +3.89 pp [-6.67, +13.89]; an alternate placeholder “[EMPTY SLOT]” yields +5.00 pp [-3.90, +14.44]. Paired differences versus primary include zero (pos2-primary -3.61 [-11.67, +4.17]; altp-h-primary -2.50 [-10.28, +5.56]). Directional estimates remain positive; this does not establish equivalence, so $C1_{\text{slot}}$ is not uniquely neutral.

L Exploratory Candidate-Conditioned Rep

Using frozen candidate correctness labels on the matched FRAMES panels, define candidate-wise $r_i = \text{Correct}(C3b_i) - \text{Correct}(C1_{\text{slot},i})$. Item-clustered means (pp; $B=2000$, seed 2026; CI withheld if fewer than 20 items): original correct +2.38 [0.00, +7.53] (73 items / 126 obs); original incorrect -8.65 [-13.30, -5.74] (183 / 474); stronger correct +5.81 [+2.28, +10.84] (146 / 327); stronger incorrect +0.37 [-2.84, +6.05] (135 / 273). Correct-candidate counts also shift between panels (original mass on 0 correct; stronger mass on 2-3). These patterns are descriptive associations under a post-generation covariate, not an identified mediator.

M Reproducibility Notes

Verified from recorded configurations and execution artifacts (no new API calls). Public OpenRouter slugs are listed below; vendor-internal build IDs are not exposed by the provider.

Primary aggregator (Luna)

openai/gpt-5.6-luna; OpenAI pin via OpenRouter; temperature 0.0; max_tokens 1024 (primary) / 4096 (matched extensions); seed 2026 where specified; request reasoning=null (effort not explicitly set; not claimed disabled).

Qwen aggregator

qwen/qwen3-32b; DeepInfra route; thinking ON; temperature 0.6; top_p 0.95; top_k 20; max_tokens 8192 in the wire specification; candidate content fixed.

Semantic judge

qwen/qwen3-32b as a separate role; temperature 0.0; max_tokens 256; seed 2026; YES/NO prompt; not the Qwen aggregator configuration.

Original candidates

upstage/solar-mini4, qwen/qwen3.8-flash,
anthropic/claude-haiku-4.5; temperature 0.0;

max_tokens 1024; sample seeds 2026–2028 where applicable; no explicit reasoning-effort override.

Stronger candidates

google/gemini-3.8-flash, x-ai/grok-4.7,
anthropic/claude-sonnet-5.5; temperature 0.0;
reasoning effort high; max_tokens 65536; seed 2026;
frozen provider routing.

Additional verified details:

- Panel sizes: $k=3$ for sampling/evidence/matched stronger aggregation; pairwise $k=2$ for highlighted model-diversity tables.
- Aggregation ceilings: 1024 for many primary runs; 4096 for the evidence ceiling check and matched stronger-panel extension (aggregation stage, not stronger candidate generation).
- Slot placeholder (primary): “(No answer was provided for this slot.)”; a sensitivity uses “[EMPTY SLOT]” and an alternate fill position.
- All candidate-aware aggregation conditions share one fixed aggregator template; only the candidate-slot contents differ.
- Bootstrap: paired complete-case, $B=2000$, seed 2026 where specified.
- Scoring: exact/normalized match, then frozen semantic judge; human audit stratified $n=60$ as above. Truncation finish-reason and token-ceiling flags are retained; the matched stronger complement recorded truncations but kept the matched denominator. Remaining unresolved: provider-default reasoning behavior when effort is unspecified and prospective provider pins never set on some original candidate calls.