ADMISSION BEFORE ALLOCATION: EXACT OPERATOR-ORDER COUNTERFACTUALS FOR MULTIMODAL MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

A stale claim can be rejected by a support verifier and still alter an answer: if rejection occurs after a finite citation slate is packed, the vacancy is not refilled with the eligible evidence that lost the allocation opportunity. ReCAP turns this failure into an exact fixed-snapshot intervention, comparing $S_B(A_\tau(C))$ with $A_\tau(S_B(C))$ while pairing the candidate sequence, support scores, coverage targets, prompt, random choices, generator, and 2,048-token budget. Across the full ScienceQA, DocVQA, and ChartQA test sets, admission before allocation yields a macro task-score gain of +0.47 [0.18, 0.76] points ($p = 0.008$), lowers unsupported-answer rates by 0.79/0.77/0.69 points, and removes 0.34–0.43k median tokens from the pre-pack proposal. In five-turn memories that evolve under each serving policy, lineage-confirmed stale carry-over falls from 8.5 to 5.4% over 167 same-image ScienceQA sessions and from 11.8 to 7.6% over 500 document-disjoint MP-DocVQA sessions. With admission fixed first, the complete typed-assembly package reaches 82.1/77.6/69.3 task score and 7.31/6.58/9.48% answer-level hallucination, outperforming the tested top-100 cross-encoder package by 0.6/0.4/0.4 score points. The evidence identifies eligibility-before-allocation as a consequential serving invariant for citation-constrained multimodal memory.

1 INTRODUCTION

An invoice assistant remembers `total = $123.50`; the visible page now reads `Total: $132.50`. Suppose similarity places the old claim ahead of the correcting OCR span in a 2,048-token citation slate. A late verifier removes $123.50 after packing, but one-pass serving does not replay allocation, so $132.50 never reaches the generator. Admitting against the visible page before packing instead makes that capacity available to the correcting span.

This example exposes an ordering question that retrieval scores and post-generation factuality do not answer: should present support determine eligibility before or after scarce context is committed? Citation systems can verify a chosen source, and memory systems can mark obsolete state, but a post-selection check cannot reconstruct a bundle whose capacity was allocated under a different feasible set. The relevant technical object is therefore not another relevance score; it is the composition of support admission and finite-budget allocation.

ReCAP makes the composition experimentally identifiable. Current-evidence admission A_τ tests each claim against compatible evidence visible now, and deterministic lazy-greedy allocation S_B packs indivisible claim–citation units. The primary comparison evaluates $S_B(A_\tau(C))$ against $A_\tau(S_B(C))$ on the same fixed snapshot. Candidate sequence, calibrated grounding scores, coverage targets, tie breaks, prompt, random draws, generator, and budget are identical; only operator order changes. A separate experiment compares flat retrieval, frozen top-100-to-20 cross-encoder reranking, and bounded typed-neighborhood assembly after admission-first serving is fixed.

The paired operators fail to commute in every evaluated consequence. Their fixed-snapshot task-score difference is +0.47 [0.18, 0.76] points at the three-dataset macro level ($p = 0.008$), with utility preserved on ScienceQA and gains of 0.8 and 0.4 points on DocVQA and ChartQA. Across 10,000 adjudicated question–seed outputs per dataset, unsupported-answer risk falls by 0.69–0.79 points with Holm-corrected $p \leq 0.011$, while median pre-pack proposal demand falls by 0.34–0.43k tokens.

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.

Closed-loop reuse supplies a distinct longitudinal test: at turn 5, the admission-first policy reduces lineage-confirmed stale carry-over from 8.5 to 5.4% in the 167 five-turn ScienceQA sessions and from 11.8 to 7.6% in 500 MP-DocVQA sessions.

The primary contribution is an exact, executable operator-order counterfactual for fixed-snapshot citation allocation. It establishes that eligibility timing changes task score, unsupported-answer risk, and evidence pressure without changing the candidates or scorer. The accompanying lineage protocol then follows each recurring unsupported proposition to an earlier served memory ID and source page under independently evolving policies. Finally, the complete typed-assembly package provides a systems endpoint after order is fixed, attaining 82.1/77.6/69.3 task score and improving over the tested cross-encoder package by 0.6/0.4/0.4 points.

2 RECAP: RE-GROUNDED ADMISSION BEFORE ALLOCATION

2.1 REPRESENTATION AND CURRENT EVIDENCE

The fixed generator is Qwen2.5-VL-72B-Instruct (Bai et al., 2025). SigLIP ViT-SO400M (Zhai et al., 2023), GroundingDINO-1.6 (Liu et al., 2024b; IDEA Research, 2024), SAM 2 (Ravi et al., 2025), PaddleOCR v2.8 (Du et al., 2020; PaddlePaddle, 2024), and TrOCR (Li et al., 2023b) provide current-turn RegionNode and OCRNode evidence. These pretrained perception and generation modules are frozen. Memory stores the evidence objects, ToolNodes, and typed ClaimNodes. The modules fitted specifically for ReCAP are DeBERTa-v3-large (He et al., 2023), fine-tuned on MNLI+SNLI for claim relations, and the isotonic grounding calibrators; this distinction avoids treating the frozen pretrained stack as unlearned.

Processing follows a fixed chronology. Perception first maps the current image or page to $\mathcal{E}^t_{\text{vis}}$. Before memory retrieval, one current-input prepass of the fixed generator produces a draft from the question and current visual input alone. The typed atomic rewrite converts this draft into current claims C^t_{now} and coverage targets $\mathcal{U}$. For each question–seed pair, this prepass output and the resulting evidence, claims, targets, and memory snapshot are computed once and supplied unchanged to every paired branch. Candidate assembly then retrieves prior served claims $C^{<t}_{\text{mem}}$ and forms $C = C^t_{\text{now}} \cup C^{<t}_{\text{mem}}$.

Let $\phi_C(c)$ and $\phi_E(e)$ denote the frozen SigLIP representations of claim text and type-compatible evidence: RegionNodes use the image tower on the crop, while OCRNodes use the text tower on normalized span text. Region evidence also retains the detector label, box, and source identifier; OCR evidence retains confidence, box, and source identifier. The raw cosine score is mapped by a type-specific isotonic function $h_{t(c)}$:

$$\text{ground}(c, e) = h_{t(c)}(\cos(\phi_C(c), \phi_E(e))) .$$

The highest-scoring compatible current node is the claim's citation candidate. A served bundle is evidence-consistent when every included claim cites such a node above its type threshold.

2.2 TWO NON-COMMUTING SERVING OPERATORS

Current-evidence admission is

$$A_\tau(C) = \left\{ c \in C : \max_{e \in \mathcal{E}^t_{\text{vis}}(c)} \text{ground}(c, e) \geq \tau_{t(c)} \right\} .$$

Here $\mathcal{E}^t_{\text{vis}}(c)$ contains only current nodes compatible with the claim type. Grounding scores are isotonic-calibrated (Guo et al., 2017) on a disjoint 20% calibration split. Headline results use each dataset's calibrated per-type operating point; the frozen-transfer study instead carries source thresholds unchanged. The ScienceQA thresholds are 0.62/0.68/0.71/0.58 for presence/attribute/count/OCR claims.

Budgeted selection applies deterministic lazy-greedy coverage (Lin & Bilmes, 2010) to indivisible claim–citation units. Let $\ell(c)$ be the exact generator-tokenizer cost of a claim and its attached citation. The indicator $g(c, u)$ equals one when c and target u have the same type and either identical normalized values or token-set Jaccard similarity above 0.85. Reported runs use uniform target weights. Define

$$F(D;\mathcal{U}) = \sum_{u \in \mathcal{U}} \mathbb{1}[\exists c \in D : g(c, u) = 1], \qquad S_B(C;\mathcal{U}) = \text{Pack}_B(\text{LG}(C, F)) , \quad B = 2048.$$

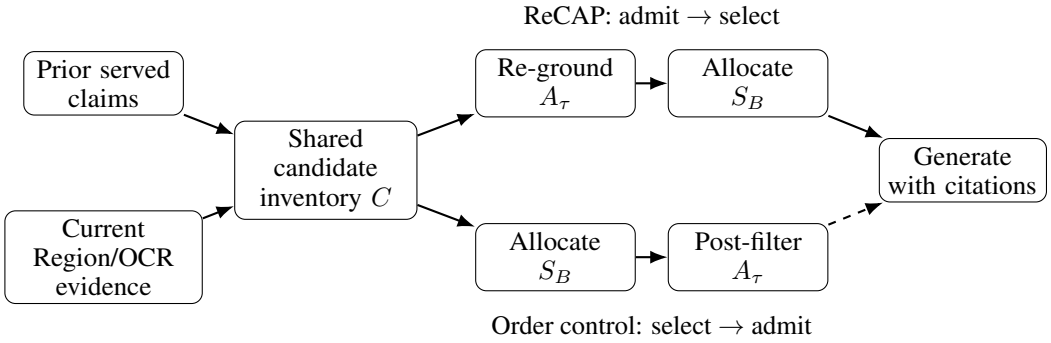

Figure 1: The fixed-snapshot intervention swaps only admission and allocation. Admit-then-select constructs the slate over eligible evidence; select-then-admit can remove an ineligible packed item without reconstructing the slate.

Starting from an empty ordered slate, LG repeatedly appends the remaining unit with greatest marginal uncovered-target gain per token, recomputes gains after each append, and stops when no positive gain remains. Deterministic tie breaks use current grounding and then claim ID. Pack_B scans the complete slate, appends a unit when it fits, skips it otherwise, and continues scanning. Thus every S_B output satisfies $\sum_{c \in S_B} \ell(c) \le B$. Pre-pack demand is $\sum_{c \in \mathrm{LG}} \ell(c)$ and may exceed B; served tokens are measured after hard packing and, for Select→Admit, after post-filtering. The atomic target set is produced once by the current-input prepass using a fixed typed rewrite (Min et al., 2023); the same 0.85 rule removes duplicate claims.

ReCAP executes $S_B(A_\tau(C); \mathcal{U})$. The exact order control executes $A_\tau(S_B(C; \mathcal{U}))$. Both use the same flat candidate pool, grounding scores, targets, allocator, prompt, budget, seeds, and tie breaks. Their bundle-level order-difference set is

$$E_{\mathrm{early}} = S_B(A_\tau(C); \mathcal{U}), \qquad E_{\mathrm{late}} = A_\tau(S_B(C; \mathcal{U})).$$

The set difference $E_{\mathrm{early}} \setminus E_{\mathrm{late}}$ contains eligible units that reach the admit-first bundle but not the post-filtered bundle. Because lazy-greedy gains are recomputed after every append, this difference records the complete consequence of changing the feasible set before selection; it need not be attributed to token capacity alone. A direct capacity-blocking witness is the concrete special case in which an inadmissible packed unit prevents a later eligible unit from fitting, and post-filtering leaves the resulting vacancy unused.

The control represents one-pass post-filtering and isolates the operator-order intervention on a fixed snapshot. Restarting the deterministic allocator over the admitted pool computes $S_B(A_\tau(C); \mathcal{U})$ and therefore collapses to the ReCAP branch by definition. The operational Verifier-gated baseline supplies a separate late-verification reference, while candidate assembly is held constant for the exact order test.

2.3 Candidate assembly variants

ReCAP-Flat retrieves the top 20 prior claims by cosine similarity and adds current claims. ReCAP-Flat-XRerank retrieves a top-100 flat pool, scores query–claim pairs with the frozen 568M-parameter `BAAI/bge-reranker-v2-m3` cross-encoder (Chen et al., 2024a; Beijing Academy of Artificial Intelligence, 2024), retains 20, and uses no relation or graph-derived feature. ReCAP-Graph starts from the same flat top-20 seeds and applies one bounded typed-neighborhood expansion, capped at 5 neighbors per edge type. Its ranking score is

$$s(c) = \cos(\mathbf{q}, \mathbf{c}) - 0.3\mathbb{1}[\mathcal{N}_{\texttt{contr}}(c) \neq \emptyset],$$

with a one-time contradiction penalty and ties resolved by maximum current grounding. `supports` and `contradicts` are directed claim-to-claim relations instantiated above the fixed NLI threshold and traversed in either incident direction for retrieval. A `derived_from` edge points from a claim to its originating RegionNode, OCRNode, or ToolNode; for expansion, an inverse provenance index maps that source node to other claims derived from it. The claim–source–claim lookup counts as one

Algorithm 1 Candidate assembly followed by the shared ReCAP contract.

Observe current evidence $\mathcal{E}^t_{\text{vis}}$
Run the shared current-input prepass; cache C^t_{now} and $\mathcal{U}$
Retrieve prior seeds from the shared memory snapshot
if variant = ReCAP-Flat-XRerank **then**
    Retrieve flat top-100; cross-encode and retain 20
else if variant = ReCAP-Graph **then**
    Expand flat top-20 by one bounded typed-neighborhood step
else
    Retain the flat top-20
end if
Form $C \leftarrow C^t_{\text{now}} \cup C^{<t}_{\text{retrieved}}$
if order = select-then-admit **then**
    $E \leftarrow A_\tau(S_B(C;\mathcal{U}))$; do not refill rejected slots
else
    $E \leftarrow S_B(A_\tau(C);\mathcal{U})$
end if
Generate from current input and the cited evidence in E

Table 1: Exactness of the controlled comparisons. "Shared" denotes the same fixed inputs or code paths within a paired run.

Component	*Order ctrl.*	*Flat*	*XRerank*	*Graph*
Memory/current evidence	shared	shared	shared	shared
Grounding, targets, selector	shared	shared	shared	shared
Prompt, budget, seeds	shared	shared	shared	shared
Initial candidate retrieval	top-20	top-20	top-100	top-20
Candidate expansion	none	none	cross-encoder	typed neighborhood
Operator order	$S_B \to A_\tau$	$A_\tau \to S_B$	$A_\tau \to S_B$	$A_\tau \to S_B$

bounded provenance-neighborhood expansion, capped like the relation neighborhoods. Duplicate normalized endpoints are merged, provenance is retained with the served claim, and current-evidence grounding is recomputed each turn. These edges affect assembly only; admission, target construction, allocation, generation, and evaluation remain shared.

2.4 Two controlled estimands

The design separates a fixed-snapshot order effect from an assembly-package effect. For any paired single-turn metric m—task score, unsupported-answer risk, or evidence demand—we report

$$\Delta^m_{\text{order}} = m(\text{ReCAP-Flat}) - m(\text{Select} \to \text{Admit}),$$
$$\Delta^m_{\text{assembly}} = m(\text{ReCAP-Graph}) - m(\text{ReCAP-Flat-XRerank}).$$

Every Δ^m_{order} pair shares the candidate sequence and all random choices, so its only intervention is whether present support is checked before or after allocation. Every $\Delta^m_{\text{assembly}}$ pair shares admission-first serving and contrasts the complete typed-neighborhood assembly with a strong graph-free top-100-to-20 cross-encoder assembly. The first estimand measures the downstream consequence of operator order under a finite budget; the second compares complete assembly packages rather than attributing their entire difference to any single edge type. Verifier-gated and the broader memory baselines provide operational reference points, while the component decomposition localizes contributions within the graph package.

The multi-turn study answers a different, longitudinal question. Both policies start each session from the same reset state, then write the claims they actually served; after the first differing bundle, their memory inventories may diverge. Turn-5 carry-over is therefore a cumulative closed-loop policy effect, not a repetition of the fixed-snapshot order estimand. Keeping these analyses separate prevents candidate-organization gains or downstream memory evolution from being attributed to the single operator swap.

3 EVALUATION

3.1 TASKS, SYSTEMS, AND COMPUTE

We evaluate ScienceQA (Lu et al., 2022), DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), and the multi-page MP-DocVQA environment (Tito et al., 2023). We use each benchmark's official test split and task-native evaluator: exact predicted-option accuracy for ScienceQA, ANLS for DocVQA, and relaxed accuracy for ChartQA, each multiplied by 100. We call these values task scores and compare methods within, not absolute levels across, datasets. Each full-test question is paired with one preassembled memory snapshot containing prior ClaimNodes and their linked evidence and tool provenance; no compared method writes to that snapshot during the paired run. Full-test results therefore estimate the conditional order effect under matched memory. Every paired method receives the same snapshot, current-input prepass, final prompt, decoding configuration, and random draws.

All primary runs use Qwen2.5-VL-72B-Instruct (Bai et al., 2025) with the fixed perception stack above, CUDA 12.4, PyTorch 2.3.1, and vLLM tensor parallelism on 4× NVIDIA A100 80GB GPUs. Results are mean±standard deviation over seeds 13, 37, 61, 71, and 97. Text memory, Citation-only, Region memory, matched-harness ALCE-style (Gao et al., 2023) and RA-CM3-style (Yasunaga et al., 2023) rerankers, and Verifier-gated provide the broader reference surface in the appendix. Verifier-gated applies its support check after a tentative bundle is formed; the exact Select→Admit row is the causal control because it reuses ReCAP-Flat's candidate order and admission scores unchanged.

3.2 METRICS AND UNCERTAINTY

Full-test task score uses questions as units and reports seed variation. Answer-level hallucination is the fraction of generated answers containing at least one unsupported atomic claim. Served-claim evidence precision uses claims as units, and non-empty coverage uses questions. At turn t, traced carry-over pools current answer claims and counts those that are both unsupported and lineage-matched to an earlier served memory ID; it is a claim rate, not an answer rate. A response with no extracted atomic claims contributes zero to both pooled counts rather than an undefined per-response ratio. The expanded reliability audit contains 2,000 adjudicated questions per dataset—the original fixed 500 plus 1,500 disjoint examples—and 10,000 question–seed outputs per method, all disjoint from calibration. Three independent annotators reach Fleiss' $\kappa = 0.76/0.72/0.68$ on ScienceQA/DocVQA/ChartQA, with disagreements adjudicated.

Task-score comparisons use paired question-level permutation tests with 10,000 permutations; a permutation swaps method labels jointly for the five paired seeds of a question. Reliability intervals and risk differences use question-cluster bootstrap resampling. Carry-over intervals resample original sessions and keep all five seeded trajectories for a sampled session together. We report two-sided 95% intervals, a pre-specified three-dataset macro-average, and Holm correction across the three per-dataset comparisons within each metric. Pre-pack evidence tokens measure the complete positive-gain lazy-greedy proposal before Pack_B; served tokens measure the bundle after tokenizer-exact hard packing and any post-filter. Both exclude the question and system prompt. Retrieval latency includes indexing, admission, candidate assembly, and allocation but excludes perception and generator calls; end-to-end latency includes the complete pipeline.

3.3 TWO DIRECT REUSE ENVIRONMENTS

The ScienceQA protocol draws 500 same-image sessions with 3–5 distinct questions per image: 166/167/167 sessions have 3/4/5 turns. Memory resets between sessions, questions are randomly ordered without replacement, and current perception remains available at every turn. Turn-5 rates use the 167 five-turn sessions; session-end rates use all 500.

The second environment draws 500 five-turn sessions from the official held-out MP-DocVQA split, one document per session, requiring at least five distinct questions. Calibration and evaluation documents are disjoint. A fresh current page is presented at each turn, question order is randomized without replacement, and prior claims retain their source page and serving-turn identifiers. A current error links to prior carry-over only when its typed normalized proposition matches an earlier

Table 2: Matched-budget results with Qwen2.5-VL-72B-Instruct and $B = 2048$. Task score is full-test mean±standard deviation over five seeds: ScienceQA exact-choice accuracy, DocVQA ANLS, and ChartQA relaxed accuracy, all on a 0–100 scale. Hallucination is the adjudicated answer-level rate over 10,000 outputs per dataset. Tokens are median pre-pack proposal demand in thousands; every served bundle obeys B. S/D/C denotes ScienceQA/DocVQA/ChartQA.

Method	*ScienceQA*	*DocVQA*	*ChartQA*	*Halluc. S/D/C*	*Tok. S/D/C*
Verifier-gated	81.0±0.8	76.1±0.9	67.9±1.0	8.72/7.46/10.53	1.52/1.86/1.72
Select → Admit	80.7±0.9	75.9±1.0	67.8±1.1	8.91/7.68/10.76	1.46/1.79/1.66
ReCAP-Flat	80.9±0.8	76.7±0.9	68.2±1.0	8.12/6.91/10.07	1.12/1.36/1.28
ReCAP-Flat-XRerank	81.5±0.8	77.2±0.9	68.9±1.0	7.79/6.75/9.91	1.10/1.34/1.25
ReCAP-Graph	82.1±0.8	77.6±0.9	69.3±1.1	7.31/6.58/9.48	1.08/1.32/1.24

unsupported served claim and its retrieval lineage contains that memory ID. For session s, if $H_{s,t}$ is the set of such traced errors at turn t, amplification is

$$\mathrm{Amp} = \frac{\sum_s \sum_{t=2}^{5} |H_{s,t} \cap H_{s,1}|}{\sum_s |H_{s,1}|}.$$

We also report persistence AUC over turns 2–5 and interference after 0/1/2 inserted irrelevant queries.

4 Results

4.1 Operator order changes what reaches the generator

The exact order pair provides the fixed-snapshot counterfactual. ReCAP-Flat changes only $S_B \to A_\tau$ to $A_\tau \to S_B$ and changes task score by +0.2/+0.8/+0.4 points, for a macro paired difference of +0.47 [0.18, 0.76] ($p = 0.008$). ScienceQA is a preserved-utility result—80.9±0.8 versus 80.7±0.9—accompanied by lower unsupported-answer risk and evidence demand. Across S/D/C, hallucination risk differences are $-0.79/-0.77/-0.69$ points with Holm-corrected $p = 0.003/0.006/0.011$, while median pre-pack proposals shrink by 0.34–0.43k tokens. Raw counts and clustered intervals appear in Appendix Table 10.

The complete typed-assembly package improves all three axes after admission order is fixed. Compared with the tested XRerank package, it gains 0.6/0.4/0.4 task-score points, with a macro difference of +0.47 [0.19, 0.74] ($p = 0.004$), while observed unsupported-answer rates fall by 0.48/0.17/0.43 points. Relative to Verifier-gated, its full-test gain is +1.1/+1.5/+1.4, with aggregate $\Delta = 1.33$ [0.81, 1.86] and $p = 0.006$. This is a complete-package comparison: Graph and XRerank intentionally expose different candidate sets, whereas the order pair above does not.

4.2 Admission-first policies suppress multi-turn carry-over

The longitudinal ordering replicates across visually and operationally distinct settings. On same-image ScienceQA, Select→Admit and ReCAP-Flat end at 8.5 [7.6,9.4]% and 5.4 [4.7,6.1]% traced carry-over; the complete typed package reaches 4.1%, while amplification moves from 2.2 to 1.4 to 1.1 linked repeats per initial error. On MP-DocVQA, Select→Admit and ReCAP-Flat end at 11.8 [10.7,12.9]% and 7.6 [6.7,8.5]%; XRerank and ReCAP-Graph reach 6.4% and 5.2%. Once each branch writes its own served claims, Select→Admit versus ReCAP-Flat is a cumulative comparison between the two serving policies; the XRerank and Graph endpoints additionally include their assembly choices. Each traced error still identifies its originating claim and source page, distinguishing stale re-entry from mere answer similarity.

4.3 Structure, robustness, and the systems frontier

The component decomposition in Appendix Table 11 preserves the same pattern on all three datasets. Proximity supplies the largest observed task-score step; contradiction scoring trades 0.1 ChartQA point for a 0.1 point hallucination reduction. Adding `derived_from` provenance then reaches 69.3

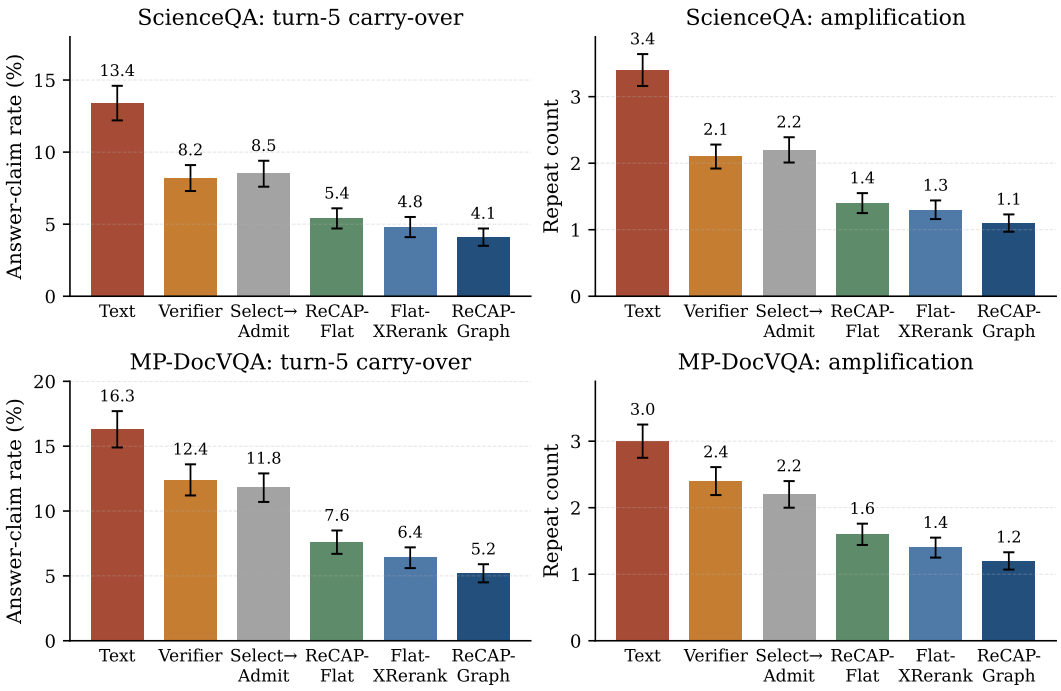

Figure 2: Cumulative policy effects on traced stale-claim carry-over and amplification in two reuse environments. Carry-over is the pooled fraction of current answer claims that are unsupported and lineage-linked to earlier memory. Bars show five-seed means; error bars are session-cluster 95% intervals. ScienceQA uses 167 five-turn sessions at turn 5, and MP-DocVQA uses 500 document-disjoint five-turn sessions.

ChartQA score and 9.6% hallucination on the original 500-question replication block; the expanded 2,000-question estimate is 9.48%. Deterministic evidence-derived targets and direct-claim targets retain the Graph-over-Flat ordering, showing that the result is stable to how the shared coverage target is constructed.

Admission first also changes the budget tail. In the exact order pair, Select→Admit has p95 pre-pack demand of 2.74/3.18/2.98k and 21.4/36.9/30.1% of S/D/C proposals above B; ReCAP-Flat reduces these to 2.09/2.47/2.34k and 6.8/12.7/10.6%. ReCAP-Graph reaches 1.96/2.29/2.18k and 3.9/8.6/7.2%. These tails explain why hard packing remains active even when Table 2's medians fall below 2,048. Typed expansion adds retrieval work, but the shorter served context keeps the end-to-end latency frontier competitive.

The calibrated ScienceQA operating points span 86.2–92.8% evidence precision and 89.4–95.6% non-empty coverage across claim types. Carrying thresholds unchanged across datasets or backbones costs 0.3–0.5 task-score point relative to target-specific thresholds. Alternative target constructions preserve the Graph-over-Flat ordering, and the appendix reports the corresponding calibration, backbone, latency, and contamination analyses.

5 Related Work

ALCE attaches and evaluates citations during generation, RA-CM3 retrieves multimodal demonstrations, and LLatrieval verifies and reranks already retrieved passages (Gao et al., 2023; Yasunaga et al., 2023; Li et al., 2024b). Their scoring improvements act on evidence that has entered the retrieval or generation path. MCiteBench evaluates multimodal citation quality (Hu et al., 2025); MCite-RL and CoRe-MMRAG improve visual search and cross-source reconciliation (Zhao et al., 2026; Tian et al., 2025). ReCAP tests a different operator location: whether a prior claim proves support against the evidence visible now before, rather than after, finite citation capacity is assigned.

RAG and multimodal generators optimize retrieval, representation, or generation quality (Lewis et al., 2020; Borgeaud et al., 2022; Izacard et al., 2023; Asai et al., 2024; Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2024a). Output audits measure factuality after generation (Min et al., 2023; Li et al., 2023c; Chen et al., 2023; Guan et al., 2024). In contrast, ReCAP tests whether unsupported content re-enters the evidence path before generation and compares the resulting bundles under an exact order swap.

Agent-memory systems and benchmarks study retention, replay, updating, and long-horizon state (Packer et al., 2023; Yao et al., 2023; Schick et al., 2023; Shinn et al., 2023; Park et al., 2023). STALE and MemOps make obsolete state and lifecycle operations explicit (Chao et al., 2026; Hao et al., 2026), but do not pair pre-allocation admission with its post-allocation order counterfactual under an identical candidate sequence. ReCAP adds that controlled comparison and follows any surviving error back to the earlier served claim and source page. Table 3 summarizes the resulting operator-level distinction.

Table 3: Decision-boundary comparison. CE re-grounding recomputes support against current evidence; pre-budget admission precedes context allocation; traced carry-over links a current error to an earlier served claim.

Work	*CE ground*	*Pre-budget*	*Fixed B*	*Trace*	*Order ctrl.*	*Flat ctrl.*	*Multi-env.*
ALCE	no	no	partial	no	no	no	no
RA-CM3	no	no	no	no	no	no	no
LLatrieval	partial	no	no	no	no	no	no
MCiteBench	no	no	no	no	no	no	no
MCite-RL	partial	no	no	no	no	no	no
CoRe-MMRAG	partial	no	no	no	no	no	no
STALE	partial	no	no	no	no	no	no
MemOps	partial	no	no	no	no	no	no
ReCAP	yes	yes	yes	yes	yes	yes	yes

Citation systems establish source support, and memory benchmarks expose obsolete state or lifecycle failures. ReCAP joins these concerns at the serving boundary and uses matched operator order to identify the consequence of allocating citation context before eligibility is known.

6 Discussion

The results separate a serving invariant from an assembly choice. Flat admission-first serving captures the exact order effect with little retrieval overhead when cosine search already exposes the relevant claim. Cross-encoding strengthens ambiguous query–claim matches, while typed neighborhoods add contradiction and evidence-ancestry routes; the complete typed package gains 0.6/0.4/0.4 task-score points over the tested XRerank package and occupies a competitive end-to-end latency frontier. A system can therefore enforce eligibility before allocation independently of whether it adopts the richer assembly.

The evidence establishes this conclusion for inspectable current evidence, calibrated claim-type thresholds, and a fixed citation budget across repeated-image and document-disjoint changing-page sessions. Empty bundles fall back to the current input. Historical or cross-page facts without currently observable support require an expanded evidence interface, and each new evidence type requires corresponding calibration. Graph versus XRerank is intentionally a complete-package comparison with different candidate exposure; the component study identifies where gains enter but is not an edge-only cardinality-matched contrast. The exact fixed-snapshot estimand, the cumulative online endpoints, and the cross-backbone package replication consequently answer complementary questions without conflating their intervention scopes.

7 Conclusion

Finite citation context makes eligibility timing a serving invariant. Under matched per-question memory snapshots, ReCAP isolates that invariant with an exact operator swap: admission before

allocation preserves or improves task score across three benchmarks while lowering unsupported-answer risk and proposal demand. Independently evolving five-turn memories show the longitudinal consequence, reducing lineage-confirmed carry-over from 8.5 to 5.4% on same-image ScienceQA and from 11.8 to 7.6% on changing-page MP-DocVQA. After the order decision is fixed, the complete typed-assembly package further improves over the tested cross-encoder package. The resulting systems rule is direct: establish present support before scarce citation capacity is committed.

Ethics Statement

Three annotators evaluated generated outputs and visible cited evidence from public benchmarks under the written support-labeling instructions in the appendix. No annotator attributes or behavioral measurements were collected, and the annotation interface contained no private user data. Disagreements were adjudicated using the same evidence display and label taxonomy.

Generative-Tool Disclosure

Generative tools assisted with manuscript language, presentation, and formatting/reference consistency during submission preparation.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. doi: 10.52202/068431-1723. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Beijing Academy of Artificial Intelligence. BAAI/bge-reranker-v2-m3 model card, 2024. URL https://huggingface.co/BAAI/bge-reranker-v2-m3.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pp. 2206–2240, 2022. URL https://proceedings.mlr.press/v162/borgeaud22a.html.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. STALE: Can LLM agents know when their memories are no longer valid? *arXiv preprint arXiv:2605.06527*, 2026. URL https://arxiv.org/abs/2605.06527.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024a. URL https://arxiv.org/abs/2402.03216.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, et al. FELM: Benchmarking factuality evaluation of large language models. *Advances in Neural Information Processing Systems*, 2023. doi: 10.52202/075280-1927. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8b8a7960d343e023a6a0afe37eee6022-Abstract-Datasets_and_Benchmarks.html.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b. URL https://arxiv.org/abs/2412.05271.

Yuning Du, Chenxia Li, Ruoyu Guo, et al. PP-OCR: A practical ultra lightweight OCR system. *arXiv preprint arXiv:2009.09941*, 2020. URL https://arxiv.org/abs/2009.09941.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, et al. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Xixuan Hao, Zeyu Zhang, Zehao Lin, Yihang Sun, Ziliang Guo, Xichong Zhang, Yuxuan Liang, Feiyu Xiong, and Zhiyu Li. MemOps: Benchmarking lifecycle memory operations in long-horizon conversations. *arXiv preprint arXiv:2607.12893*, 2026. URL https://arxiv.org/abs/2607.12893.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

Caiyu Hu, Yikai Zhang, Tinghui Zhu, Yiwei Ye, and Yanghua Xiao. MCiteBench: A multimodal benchmark for generating text with citations. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 5949–5966. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.318. URL https://aclanthology.org/2025.findings-emnlp.318/.

IDEA Research. Grounding DINO 1.6 release notes. Official release notes in the Grounding DINO API repository, 2024. URL https://github.com/IDEA-Research/Grounding-DINO-1.5-API.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023. URL https://www.jmlr.org/papers/v24/23-0037.html.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a. URL https://arxiv.org/abs/2408.03326.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 19730–19742, 2023a. URL https://proceedings.mlr.press/v202/li23q.html.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. TrOCR: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13094–13102, 2023b. doi: 10.1609/aaai.v37i11.26538. URL https://ojs.aaai.org/index.php/AAAI/article/view/26538.

Xiaonan Li, Changtai Zhu, Linyang Li, Zhangyue Yin, Tianxiang Sun, and Xipeng Qiu. LLatrieval: LLM-verified retrieval for verifiable generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5453–5471. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.naacl-long.305. URL https://aclanthology.org/2024.naacl-long.305/.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023c. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Hui Lin and Jeff Bilmes. Multi-document summarization via budgeted maximization of submodular functions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 912–920, 2010. URL https://aclanthology.org/N10-1134/.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.html.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, 2024b. doi: 10.1007/978-3-031-72970-6_3. URL https://mlanthology.org/eccv/2024/liu2024eccv-grounding/.

Pan Lu, Swaroop Mishra, Tony Xia, et al. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2209.09513.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Mathew_DocVQA_A_Dataset_for_VQA_on_Document_Images_WACV_2021_paper.html.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, et al. 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*, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. URL https://arxiv.org/abs/2310.08560.

PaddlePaddle. PaddleOCR v2.8.0. Official software release, 2024. URL https://github.com/PaddlePaddle/PaddleOCR/releases/tag/v2.8.0.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023. doi: 10.1145/3586183.3606763. URL https://arxiv.org/abs/2304.03442.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, et al. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Ha0qR7LsyE.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, et al. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2302.04761.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2303.11366.

Yang Tian, Fan Liu, Jingyuan Zhang, Victoria W., Yupeng Hu, and Liqiang Nie. CoRe-MMRAG: Cross-source knowledge reconciliation for multimodal RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32967–32982. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1583. URL https://aclanthology.org/2025.acl-long.1583/.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multi-page DocVQA. *Pattern Recognition*, 144:109834, 2023. doi: 10.1016/j.patcog.2023.109834. URL https://doi.org/10.1016/j.patcog.2023.109834.

Shunyu Yao, Jeffrey Zhao, Dian Yu, et al. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, et al. Retrieval-augmented multimodal language modeling. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. URL https://proceedings.mlr.press/v202/yasunaga23a.html.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14433–14443, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

Suifeng Zhao, Zida Liu, Xinyu Lei, Lei Sun, Jun Gao, and Sujian Li. MCite-RL: Towards reliable multimodal RAG via citation-enhanced agentic reinforcement learning. *arXiv preprint arXiv:2608.21808*, 2026. URL https://arxiv.org/abs/2608.21808.

A Implementation and Evaluation Details

A.1 Atomic claims and coverage targets

The atomic-claim rewriter uses one fixed prompt across datasets and backbones. The draft is generated once from the current question and visual input, without retrieved memory or selected citations, before any compared method runs. Its output is cached by question–seed pair. The rewriter emits a JSON array with fields `type`, `text`, and optional `normalize`; the same array supplies C^t_{now} and the unit-weight coverage targets $\mathcal{U}$ to every paired method. Allowed types are `presence`, `attribute`, `count`, and `ocr`. Conjunctions are split, numbers and currencies are normalized, and malformed outputs are rejected rather than repaired with method-specific context.

System: You are an information extraction model. Rewrite the given draft answer into a concise set of typed atomic claims that can be independently verified by citations. Output valid JSON only.
User:

> Question: `{QUESTION}`
> Draft answer: `{DRAFT_ANSWER}`
> Task: Output a JSON array. Each element must contain:
>
> - `"type"`: one of `"presence"`, `"attribute"`, `"count"`, or `"ocr"`.
> - `"text"`: a single atomic claim with one fact.
> - `"normalize"` (optional): a normalized value for numbers, currencies, or OCR strings.
>
> Rules: split conjunctions, do not invent facts not present in the draft answer, and prefer `ocr` when a claim depends on an exact string or number from the image or document.

Final generation receives the current visual input, question, and hard-packed list of claim–citation units. The shared instruction asks the generator to answer the question using the visible input and to attach citations only from that list; the template and decoding configuration are unchanged across each paired comparison.

Table 4: Default retrieval and implementation settings used in all reported runs.

Component	*Setting*	*Role*
Flat seed retrieval	$k = 20$	ClaimNodes ranked by cosine similarity.
XRerank retrieval	top-100 $\rightarrow$ 20	Frozen BGE cross-encoder; no typed or graph-derived features.
Graph proximity	1 expansion, cap 5/type	Incident claim relations plus indexed claim–source–claim provenance neighbors.
Contradiction penalty	$\gamma = 0.3$	Applied once when a retrieved contradiction neighbor exists.
Tie break	$\max_e \text{ground}(c, e)$	Favors the claim with stronger present-evidence support.
Admission	per-type $\tau_{t(c)}$	Isotonic calibration on a disjoint held-out 20% split.
Coverage selection	lazy greedy, $B = 2048$	Deterministic proposal followed by tokenizer-exact hard packing.
Deduplication	Jaccard 0.85	Removes near-duplicate normalized claims.
OCR refinement	TrOCR below 0.7	Refines low-confidence PaddleOCR spans.
Index/store	FAISS HNSW + RocksDB	HNSW $M = 32$, efSearch 128; persistent typed metadata.

Table 5: Models and system components in the shared evaluation stack.

Component	*Model or version*	*Learned*	*Use*
Generator	Qwen2.5-VL-72B-Instruct (Bai et al., 2025)	No	vLLM tensor parallelism across four GPUs.
Claim/region encoder	SigLIP ViT-SO400M	No	Frozen text/image towers for claims, OCR spans, and region crops.
Detector/segmenter	GroundingDINO-1.6 + SAM 2	No	Current-turn proposals and masks.
OCR	PaddleOCR v2.8 + TrOCR	No	Text spans and low-confidence refinement.
Flat reranker	BAAI/bge-reranker-v2-m3 (Chen et al., 2024a; Beijing Academy of Artificial Intelligence, 2024)	No	568M cross-encoder over query–claim pairs.
Contradiction model	DeBERTa-v3-large	Yes	MNLI+SNLI fine-tuning for typed edges.
Grounding calibration	Isotonic regression	Yes	Per-type admission thresholds.
Index/store	FAISS HNSW + RocksDB	No	Candidate retrieval and typed metadata.
Selector	Lazy-greedy coverage	No	Shared budget allocation after admission.

A.2 Calibration and transfer

Table 6: ScienceQA admission calibration. Precision is evaluated at the selected threshold; coverage is the non-empty eligible-bundle rate.

Type	τ	*Precision*	*Coverage*	*ECE*
Presence	0.62	91.4%	94.2%	0.038
Attribute	0.68	88.6%	91.8%	0.046
Count	0.71	86.2%	89.4%	0.054
OCR	0.58	92.8%	95.6%	0.032
Global threshold	0.65	87.4%	91.2%	0.058

Table 7: Frozen-threshold transfer. Task-score deltas are relative to thresholds tuned on the named target, so the table measures portability of the admission calibration.

Source → target	*Precision*	*Coverage*	*Task-score* Δ
ScienceQA→DocVQA	87.8±0.6	91.4±0.5	−0.4
ScienceQA→ChartQA	86.4±0.8	90.2±0.6	−0.5
DocVQA→ScienceQA	88.1±0.5	92.0±0.4	−0.3
DocVQA→ChartQA	87.2±0.7	90.8±0.5	−0.4
Qwen→InternVL	87.6±0.6	91.2±0.5	−0.4
Qwen→LLaVA	87.1±0.7	90.6±0.6	−0.5

Headline endpoints use the target dataset's selected per-type thresholds; Table 7 is the separate source-frozen intervention. Calibration precision conditions on audited claim–evidence pairs at each type-specific operating point. The served-claim shares in Table 21 instead condition on all claims produced after admission, so the two tables summarize different populations.

B Expanded Reliability and Mechanism Results

B.1 Full matched-harness baseline surface

Text memory cosine-retrieves serialized prior ClaimNodes; Citation-only retrieves the associated source units without reusable claims; and Region memory retrieves prior RegionNodes and OCRNodes. The ALCE-style adaptation applies citation-aware reranking after flat retrieval, while the RA-CM3-style adaptation uses multimodal retrieval and reranking in the same generator harness. Verifier-gated forms a tentative evidence slate before applying its support verifier. These operational references share the memory snapshot, current input, hard budget, generator, final prompt, and evaluation with ReCAP; they are matched-harness adaptations rather than claims of bit-identical reproduction of each original system.

Table 8: Operational comparator contracts. "Late, no refill" means that rejected units leave their positions vacant; all rows share the final prompt, generator, tokenizer-exact hard budget, seeds, and evaluation examples.

Method	*Candidate object and ranking*	*Support treatment*	*Bundle construction*
Text memory	Serialized prior ClaimNodes; query–claim cosine	none before serving	shared coverage pack
Citation-only	Sources linked to cosine-retrieved claims; claim score retained	source units only	shared coverage pack
Region memory	Prior RegionNodes/OCRNodes; type-compatible cosine	none before serving	shared coverage pack
ALCE-style	Flat retrieved units; citation-aware reranking	reranker score	shared coverage pack
RA-CM3-style	Multimodal units; multimodal retrieval and reranking	reranker score	shared coverage pack
Verifier-gated	Tentative claim–citation slate	late, no refill	verify after tentative pack
Select→Admit	ReCAP-Flat's identical C, scores, and order	A_τ late, no refill	$A_\tau(S_B(C))$
ReCAP-Flat	Same flat C and scores as the order control	A_τ before packing	$S_B(A_\tau(C))$

Table 9: Original 500-question replication block. Task scores use the benchmark-native evaluators defined in the main paper. Hallucination rates remain the backed fixed-subset values; the primary 2,000-question analysis appears in Table 10.

Method	*ScienceQA*	*DocVQA*	*ChartQA*	*Halluc. S/D/C*	*Pre-pack tok. S/D/C*
Text memory	78.4±0.9	72.1±1.1	65.6±1.2	12.8/10.4/13.9	1.96/2.34/2.14
Citation-only	79.6±0.9	74.2±1.0	66.8±1.1	10.6/8.8/12.2	1.84/2.18/2.02
Region memory	80.2±0.8	75.4±1.0	67.4±1.1	9.4/8.2/11.6	1.68/2.04/1.88
ALCE-style	80.6±0.8	75.3±1.0	67.2±1.1	9.8/8.0/11.4	1.72/1.94/1.86
RA-CM3-style	80.9±0.8	75.8±1.0	67.6±1.0	9.2/7.6/10.8	1.66/1.88/1.78
Verifier-gated	81.0±0.8	76.1±0.9	67.9±1.0	8.6/7.4/10.4	1.52/1.86/1.72
ReCAP-Flat	80.9±0.8	76.7±0.9	68.2±1.0	8.2/7.0/10.2	1.12/1.36/1.28
ReCAP-Graph	82.1±0.8	77.6±0.9	69.3±1.1	7.4/6.7/9.6	1.08/1.32/1.24

This replication block uses 500 questions and five seeds, hence 2,500 outputs per dataset. Table 10 uses the larger disjoint-augmented adjudication set and is the canonical source for the 2,000-question rates reported in the main paper.

B.2 TWO-THOUSAND-QUESTION ADJUDICATION

Table 10: Answer-level hallucination on 10,000 question–seed outputs per dataset. Each cell gives rate [question-cluster 95% CI]; raw count/10,000.

Method	*ScienceQA*	*DocVQA*	*ChartQA*
Verifier-gated	8.72 [8.19,9.29]; 872/10,000	7.46 [6.97,7.98]; 746/10,000	10.53 [9.94,11.15]; 1,053/10,000
Select → Admit	8.91 [8.37,9.48]; 891/10,000	7.68 [7.18,8.21]; 768/10,000	10.76 [10.16,11.39]; 1,076/10,000
ReCAP-Flat	8.12 [7.62,8.66]; 812/10,000	6.91 [6.44,7.41]; 691/10,000	10.07 [9.50,10.68]; 1,007/10,000
ReCAP-Flat-XRerank	7.79 [7.30,8.31]; 779/10,000	6.75 [6.29,7.24]; 675/10,000	9.91 [9.35,10.51]; 991/10,000
ReCAP-Graph	7.31 [6.85,7.80]; 731/10,000	6.58 [6.13,7.06]; 658/10,000	9.48 [8.93,10.06]; 948/10,000

Against the exact Select→Admit control, ReCAP-Flat has risk differences of $-0.79/-0.77/-0.69$ points with Holm-corrected $p = 0.003/0.006/0.011$; this is the reliability estimate that isolates operator order. For operational context, XRerank differs from that control in both order and reranking and yields $-1.12/-0.93/-0.85$ ($p < 0.001/0.003/0.005$), while ReCAP-Graph additionally changes candidate assembly and yields $-1.60/-1.10/-1.28$ ($p < 0.001/<0.001/<0.001$). ReCAP-Graph's corresponding macro operational difference is -1.33 $[-1.57, -1.08]$ points. The assembly-specific comparison remains Graph versus XRerank, as defined in the main paper.

B.3 COMPLETE STRUCTURE DECOMPOSITION

Table 11: Candidate-assembly decomposition under identical admission and coverage selection. Task score is five-seed mean±standard deviation. Hallucination uses the original five-seed 500-question block; counts are unsupported outputs over 2,500.

Assembly	*Task score S/D/C*	*Halluc. S/D/C*	*Unsupported count S/D/C*
Flat	80.9±0.8 / 76.7±0.9 / 68.2±1.0	8.2/7.0/10.2	205/175/255
+ Proximity	81.4±0.8 / 77.1±0.9 / 68.8±1.1	7.8/6.8/9.9	195/170/248
+ Contradict	81.6±0.8 / 77.2±0.9 / 68.7±1.1	7.6/6.8/9.8	190/170/245
+ `derived_from`	82.1±0.8 / 77.6±0.9 / 69.3±1.1	7.4/6.7/9.6	185/168/240

Proximity recovers semantically adjacent evidence on all three datasets. Contradiction edges suppress unsupported ChartQA claims despite a 0.1-point task-score tradeoff, and the indexed `derived_from` neighborhood then reconnects normalized claims that share source regions or OCR spans. The full row therefore measures the combined recall and provenance effect rather than attributing the endpoint to contradiction scoring alone.

B.4 COVERAGE-TARGET ROBUSTNESS

The two alternative target sources preserve the Graph-over-Flat ordering in both task score and unsupported-answer rate. Their lower extraction F1 explains the small absolute shifts, while comparable coverage and proposal demand show that the ordering is not created by a single draft-rewrite path.

Table 12: Target-construction task score and 2,000-question hallucination. Values are S/D/C triplets; candidate assembly, admission, budget, generator, and seeds are fixed.

Targets	*Assembly*	*Task score*	*Hallucination*
Model rewrite	Flat	80.9/76.7/68.2	8.12/6.91/10.07
Model rewrite	Graph	82.1/77.6/69.3	7.31/6.58/9.48
Evidence-derived	Flat	80.5/76.3/67.9	8.31/7.08/10.28
Evidence-derived	Graph	81.7/77.2/68.9	7.53/6.73/9.70
Direct claims	Flat	80.3/76.0/67.7	8.44/7.19/10.42
Direct claims	Graph	81.5/77.0/68.7	7.66/6.84/9.82

Table 13: Target-construction non-empty coverage, pre-pack proposal tokens, and target extraction F1. Values are S/D/C triplets.

Targets	*Assembly*	*Coverage*	*Tokens*	*F1*
Model rewrite	Flat	96.2/94.8/93.1	1.12/1.36/1.28	0.86/0.83/0.81
Model rewrite	Graph	96.2/94.8/93.1	1.08/1.32/1.24	0.86/0.83/0.81
Evidence-derived	Flat	95.8/94.1/92.5	1.10/1.34/1.26	0.84/0.82/0.80
Evidence-derived	Graph	96.0/94.4/92.8	1.06/1.29/1.22	0.84/0.82/0.80
Direct claims	Flat	95.3/93.7/92.0	1.08/1.31/1.24	0.82/0.79/0.77
Direct claims	Graph	95.6/94.0/92.4	1.04/1.27/1.20	0.82/0.79/0.77

C Budget Pressure and Runtime

The allocator records both the complete positive-gain lazy-greedy slate and the bundle that survives tokenizer-exact packing. Lazy greedy recomputes marginal gain after each selected unit. Hard packing then scans the resulting order, skips any unit that does not fit, and continues; it never serves more than B tokens. For Select→Admit, served-token accounting occurs after the post-filter, so rejected units can leave the final bundle under-filled. Pre-pack demand diagnoses how much positive-gain evidence competes for finite context, whereas only the final bundle reaches the generator.

Table 14: Distributional citation-budget behavior. Triplets are S/D/C. Pre-pack demand may exceed B; served tokens are means after Pack_B and any post-filter and always obey the hard budget. Exceedance gives the percent and count of proposals above B over 10,000 requests.

Method	*Pre-pack median*	*Pre-pack p95*	*Above B: %; count*	*Served mean*
Verifier-gated	1.52/1.86/1.72	2.86/3.34/3.09	23.7/39.8/32.5; 2370/3980/3250	1.43/1.69/1.61
Select → Admit	1.46/1.79/1.66	2.74/3.18/2.98	21.4/36.9/30.1; 2140/3690/3010	1.39/1.63/1.57
ReCAP-Flat	1.12/1.36/1.28	2.09/2.47/2.34	6.8/12.7/10.6; 680/1270/1060	1.08/1.28/1.22
ReCAP-Flat-XRerank	1.10/1.34/1.25	2.02/2.38/2.27	5.1/10.1/8.5; 510/1010/850	1.06/1.26/1.20
ReCAP-Graph	1.08/1.32/1.24	1.96/2.29/2.18	3.9/8.6/7.2; 390/860/720	1.04/1.24/1.19

The long-tail reduction dominates the admission and assembly overhead. ReCAP-Graph performs additional typed retrieval work relative to Verifier-gated, yet its shorter evidence context lowers median end-to-end generation time. XRerank is the costliest graph-free assembler because it cross-encodes 100 query–claim pairs. The NLI sweep below provides a sensitivity view around the adopted threshold of 0.6; the complete package retains the same task-score range while retrieval p95 varies from 174 to 210 ms.

Table 15: Runtime and peak GPU memory. Latencies are p50/p95 seconds for S/D/C; retrieval covers indexing through allocation and excludes perception and generator calls, while end-to-end covers the complete pipeline.

Method	*Retrieval p50/p95 (s)*	*End-to-end p50/p95 (s)*	*Peak GB*
Verifier-gated	.112/.148; .128/.171; .121/.162	1.84/2.63; 2.12/2.98; 1.96/2.80	73.4
Select → Admit	.118/.154; .135/.179; .129/.170	1.80/2.58; 2.06/2.91; 1.91/2.73	73.6
ReCAP-Flat	.119/.156; .137/.181; .131/.173	1.66/2.35; 1.91/2.66; 1.78/2.51	73.8
ReCAP-Flat-XRerank	.184/.236; .207/.265; .196/.251	1.74/2.43; 1.99/2.75; 1.85/2.61	75.1
ReCAP-Graph	.146/.192; .164/.218; .158/.207	1.68/2.36; 1.93/2.68; 1.80/2.53	74.6

Table 16: Threshold for instantiating support and contradiction edges.

NLI threshold	*SciQA*	*DocVQA*	*ChartQA*	*Degree*	*Retr. p95*
0.4	81.8±0.8	77.2±0.9	69.0±1.1	8.6	210 ms
0.5	82.0±0.8	77.4±0.9	69.2±1.1	7.4	198 ms
0.6	82.1±0.8	77.6±0.9	69.3±1.1	6.8	192 ms
0.7	81.6±0.9	77.2±1.0	68.9±1.2	5.2	174 ms

Table 17: Cross-backbone replication of the Verifier-gated versus complete ReCAP-Graph package comparison. Values are task-native scores; this table evaluates package transfer rather than the exact order estimand. The token column is the ScienceQA median pre-pack proposal demand.

Backbone	*Task score: Verifier-gated → ReCAP-Graph (S/D/C)*	*SciQA tokens*
Qwen2.5-VL-72B (Bai et al., 2025)	81.0→82.1 / 76.1→77.6 / 67.9→69.3	1.52→1.08
InternVL2.5-38B (Chen et al., 2024b)	79.4→80.7 / 73.3→74.9 / 63.2→64.8	1.44→1.02
LLaVA-OneVision-34B (Li et al., 2024a)	78.6→80.0 / 72.8→74.1 / 62.4→63.9	1.41→0.98

D STALE-CLAIM CARRY-OVER TRACES

D.1 SAME-IMAGE SCIENCEQA

The 500 sessions contain 166/167/167 sessions of length 3/4/5. Session-end statistics pool all sessions at their final observed turn; turn-5 statistics use only the 167 five-turn sessions. Questions never repeat within a session, memory resets between sessions, and every method reruns current perception. Policies begin from the same reset state and then retain their own served claims, so later inventories can differ. The bootstrap resamples the 167 original sessions and keeps their five seeded trajectories together; raw denominators count all turn-5 answer claims from those trajectories.

Table 18: ScienceQA five-turn endpoint. Confidence intervals use session-cluster bootstrap resampling; raw counts are traced unsupported claims over all answer claims at turn 5.

Method	*Turn-5 % [95% CI]*	*Raw claims*	*Amplification*	*Carry-over AUC*
Text memory	13.4 [12.2,14.6]	108/806	3.4	11.8
Verifier-gated	8.2 [7.3,9.1]	66/805	2.1	7.1
Select $\rightarrow$ Admit	8.5 [7.6,9.4]	68/800	2.2	7.4
ReCAP-Flat	5.4 [4.7,6.1]	43/796	1.4	4.7
ReCAP-Flat-XRerank	4.8 [4.1,5.5]	38/792	1.3	4.3
ReCAP-Graph	4.1 [3.5,4.7]	33/805	1.1	3.6

The cumulative policy comparison lowers ScienceQA carry-over from 8.5% to 5.4%. XReranking and typed provenance reduce it further, and the resulting ordering matches the fixed-snapshot reliability evidence.

D.2 FIVE-TURN MP-DOCVQA

Each of 500 official held-out documents contributes one five-turn session. Documents are disjoint from calibration, questions do not repeat, the current page changes with the question, and all 500 sessions contribute to turn-5 estimates. Every policy starts from the same reset state and then evolves its own served-claim memory. Across the five seeds, each method therefore produces 2,500 session trajectories. Session-cluster intervals resample the 500 documents and retain all seed trajectories for each sampled document.

Table 19: MP-DocVQA stale-claim carry-over. Endpoints are five-seed means; $\pm$ values are seed standard deviations, confidence intervals are session-cluster 95% intervals, and raw counts are pooled traced errors over all turn-5 answer claims.

Method	*Turn-5 %: mean$\pm$sd [CI]*	*Raw claims*	*Amplification*	*AUC*
Text memory	16.3$\pm$0.8 [14.9,17.8]	396/2,430	3.0	14.1
Verifier-gated	12.4$\pm$0.7 [11.2,13.6]	301/2,427	2.4	10.7
Select $\rightarrow$ Admit	11.8$\pm$0.7 [10.7,12.9]	286/2,424	2.2	10.3
ReCAP-Flat	7.6$\pm$0.5 [6.7,8.5]	184/2,421	1.6	6.5
ReCAP-Flat-XRerank	6.4$\pm$0.5 [5.6,7.2]	155/2,422	1.4	5.6
ReCAP-Graph	5.2$\pm$0.4 [4.5,5.9]	126/2,423	1.2	4.6

Table 20: MP-DocVQA turn-5 stale-claim carry-over after 0/1/2 irrelevant inserted turns.

Method	*0/1/2 inserted turns (%)*
Text memory	16.3/18.7/21.4
Verifier-gated	12.4/14.6/17.1
Select → Admit	11.8/13.8/16.2
ReCAP-Flat	7.6/9.1/10.9
ReCAP-Flat-XRerank	6.4/7.8/9.4
ReCAP-Graph	5.2/6.3/7.7

E Annotation and External Assets

Every audited output was independently labeled by three annotators. Disagreements were adjudicated, and *unclear* remained a separate label rather than being converted to *unsupported*. Answer-level hallucination counts an output only when at least one adjudicated claim is unsupported.

Table 21: Adjudicated label distribution for all served ReCAP-Graph claims on the original fixed subsets.

Dataset	*Supported*	*Unsupported*	*Unclear*	*Fleiss' κ*
ScienceQA (500)	86.4%	7.2%	6.4%	0.76
DocVQA (500)	83.8%	8.6%	7.6%	0.72
ChartQA (500)	81.2%	10.4%	8.4%	0.68

Annotators viewed the benchmark question, generated answer, served atomic claim, and cited Region/OCR snippets. The written rule assigned *supported* only when visible evidence directly grounded the claim, *unsupported* when evidence failed to support or contradicted it, and *unclear* when the displayed evidence was insufficient or ambiguous. Outside knowledge and plausibility judgments were disallowed. Adjudication used the same evidence display and taxonomy.

Table 22: External assets and official access terms used in the evaluation.

Asset	*Role*	*License or access terms*
ScienceQA	benchmark	CC BY-NC-SA 4.0 official repository.
DocVQA / MP-DocVQA	benchmarks	DocVQA challenge-portal access terms.
ChartQA	benchmark	GPL-3.0 official repository.
Qwen2.5-VL-72B-Instruct	generator	Apache-2.0 official model card.
InternVL2.5-38B	replication	MIT official model card.
LLaVA-OneVision family	replication	Apache-2.0 official repository.
SigLIP ViT-SO400M	region encoder	Apache-2.0 checkpoint release.
GroundingDINO-1.6 / SAM 2	perception	Apache-2.0 official repositories.
PaddleOCR v2.8 / TrOCR	OCR	Apache-2.0 / MIT official repositories.
DeBERTa-v3-large	contradiction	MIT official model card.
BAAI/bge-reranker-v2-m3	flat reranking	Apache-2.0 official model card.

F Auxiliary Contamination Evidence

These auxiliary tests probe whether the same admission mechanism remains useful when stale content is intentionally made retrieval-competitive. They are separated from the natural reuse endpoints because their threat models control which source an attacker can modify.

Table 23: Threat and evidence matrix for the auxiliary contamination checks.

Attack	*Object, capability, and trust*	*Comparator and evidence*	*Utility and failure mode*
OCR-span injection	Current OCR claim; the attacker inserts plausible text but cannot alter calibrated regions.	Verifier-gated; held-out DocVQA; Table 24.	2.1% clean FPR; failures concentrate in low-contrast spans.
Tool-output contamination	Chart claim; the attacker controls one stale tool return but not the current chart.	Verifier-gated; held-out ChartQA; Table 24.	2.4% clean FPR; failures follow ambiguous legends.
Score-matched stale claim	Prior claim; the attacker chooses the highest-similarity paraphrase but cannot alter current evidence or calibration.	Verifier-gated and both flat controls; 300 held-out sessions/dataset; Table 25.	2.8% clean FPR; failures cluster in weak region proposals.

Table 24: Backed quarantine behavior for OCR and tool contamination.

Attack	*Method*	*ASR*	*Quarantine*	*Clean FPR*
OCR injection	Verifier-gated	42.6%	18.2%	1.6%
OCR injection	ReCAP-Graph	24.2%	56.4%	2.1%
Tool contamination	Verifier-gated	34.2%	22.8%	1.3%
Tool contamination	ReCAP-Graph	18.6%	61.2%	2.4%

OCR and tool contamination produce the same direction of change: ReCAP-Graph raises quarantine while reducing attack success at a small clean false-positive cost. The score-matched attack preserves the ordering from natural reuse: the complete typed package lowers ASR from 46.7% to 22.6% while raising quarantine recall from 31.2% to 73.8%. Its 2.8% clean false-positive rate and 58 ms added retrieval time make the security–utility trade-off explicit.

Table 25: Score-matched stale-claim attack over 300 held-out sessions per dataset (900 total). ASR includes counts; latency is added retrieval time. Admission-first variants retain lower ASR after retrieval-similarity matching.

Method	*ASR; count*	*Quarantine recall*	*Clean FPR*	*Task-score* Δ	*Latency*
Verifier-gated	46.7%; 420/900	31.2%	1.8%	-0.8	+9 ms
ReCAP-Flat	31.8%; 286/900	61.7%	2.2%	-0.5	+17 ms
ReCAP-Flat-XRerank	28.9%; 260/900	66.4%	2.5%	-0.4	+74 ms
ReCAP-Graph	22.6%; 203/900	73.8%	2.8%	-0.6	+58 ms