LIVESEARCHVQA: DIAGNOSING EVIDENCE ACCESS AND USE IN MULTIMODAL SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

An agent can answer a recent visual question without searching, or retrieve the right evidence and still answer incorrectly. End-to-end accuracy alone cannot distinguish these outcomes. We introduce LIVESEARCHVQA, a renewable benchmark that connects the construction of search-demanding questions to the diagnosis of search behavior. The collection contains 6,000 questions in 30 dated builds. Evidence-first generation and repeated visual, closed-book, and supplied-evidence checks retain questions whose answers are unexpressed without evidence but recoverable with it under a fixed panel. Evaluation pairs closed-book, with-search, and oracle-evidence runs on the same items, then separates evidence availability from answer correctness. Across twelve systems, with-search accuracies that differ by less than 0.1 percentage points can still reverse in coverage and supported correctness, and raising retrieval depth increases coverage while moving accuracy in both directions. Construction records, paired conditions, and targeted interventions therefore let us ask both whether a search agent succeeds and what information supports that success. Our code is available at: https://anonymous.4open.science/r/LiveSearchVQA.

1 INTRODUCTION

Multimodal search agents combine visual understanding with web search to answer questions whose answers lie outside the input image (Jiang et al., 2024; Chang et al., 2022). We study recent, event-specific questions: the image identifies the search target, and external sources supply the requested fact. Because new events continually create new information needs, evaluation must keep testing evidence acquisition rather than familiarity with a fixed collection (Tao et al., 2026). Final-answer accuracy leaves two things unresolved: a correct answer may not have required search, and an incorrect one does not say whether the needed evidence ever reached the agent (Chen et al., 2023). We therefore develop a live setting that tests a panel-relative information gap on each question, records the support available in each run, and renews the question pool as new facts emerge.

At the question level, recency and difficulty do not by themselves make a question a search test. A recent article can restate a familiar fact (Vu et al., 2024; Kasai et al., 2023), and a question is uninformative when its reference answer is itself unsupported (Northcutt et al., 2021). Existing benchmarks therefore add source-based answerability checks and model-based difficulty filters (Zeng et al., 2026; Fu et al., 2025). We require both on the same candidate under a declared model panel: repeated closed-book failure and repeated supplied-evidence success (Figure 1a,b). This joint criterion identifies an *observed information gap* — the panel does not express the correct answer from the image and question alone, but recovers it once the recorded supporting evidence is supplied.

At the outcome level, answer correctness does not reveal whether sufficient evidence reached the agent. One failed run may never surface the required fact; another may surface it and still answer incorrectly, consistent with the difficulty of using noisy retrieved context (Pham et al., 2026). Equal answer scores can therefore correspond to different evidence states, and a correct answer need not be accompanied by observed support (Ru et al., 2024). We assess evidence availability and answer correctness jointly, against the same reference fact and time anchor that established the question (Figure 1c).

Assessing this ability on newly emerging facts requires renewal, not a one-time collection. A fixed pool drifts: later models may already answer earlier items without search, at which point those items no longer test evidence acquisition (Li et al., 2024). Collecting newer sources is not enough,

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

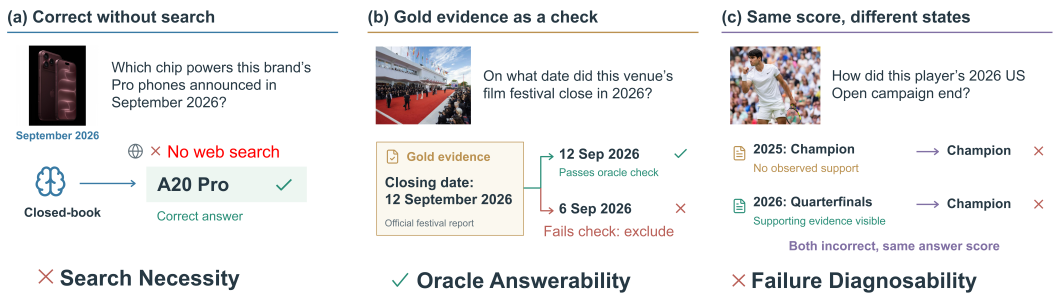

Figure 1: **Why recent questions and answer accuracy are not enough to evaluate multimodal search.** (a) *Search necessity*: a recent question can still be answered correctly without search; freshness alone does not establish a need for retrieval. (b) *Oracle answerability*: supplying gold evidence checks whether the question becomes answerable; items that fail the declared oracle criterion are excluded. (c) *Failure diagnosability*: the same incorrect answer can occur with or without sufficient observed support, leaving different limitations indistinguishable under answer accuracy.

because each new question must still elicit closed-book failure and supplied-evidence success under the same checks. We define *live renewal* as repeatedly admitting time-anchored questions under the same admission and diagnostic requirements, while preserving the answers, evidence, and evaluation records of released versions (Fu et al., 2025; Zhang et al., 2026; Kiela et al., 2021). Renewal changes which facts are tested; it does not change what the checks mean.

We instantiate this setting as LIVESEARCHVQA: 6,000 questions in 30 dated builds, each keeping a versioned reference-evidence record (Figure 7). Evidence-first generation and joint admission retain questions that exhibit closed-book failure and supplied-evidence success; fresh closed-book (CB), with-search (WS), and oracle-evidence (OR) runs then evaluate the same item under different information, and the retained record anchors support judgments on the content visible in each search trace (Figure 2). Across twelve systems, statistically indistinguishable search accuracies conceal opposite evidence profiles: GLM-5V-turbo and qwen3.7-flash differ by 0.07 points in with-search accuracy, yet reach it with coverage of 71.9% against 63.1%. Controlled interventions show the same separation, with greater evidence access improving answers for one system and degrading them for another.

Our contributions are threefold. (i) *A live evaluation setting and benchmark.* We formulate live multimodal search evaluation around newly emerging, visually grounded facts, with renewed admission and fixed reference records, and instantiate it as LIVESEARCHVQA: 6,000 questions in 30 dated builds (Table 1). (ii) *A construction-linked diagnostic protocol.* We connect repeated closed-book and supplied-evidence admission to paired CB/WS/OR evaluation, reusing the retained evidence records to judge observed support jointly with answer correctness. (iii) *Empirical insight into search behaviour.* Across twelve systems and controlled evidence interventions, similar answer scores conceal different evidence profiles, and increasing evidence coverage need not improve answer accuracy.

2 RELATED WORK

Establishing that a question tests search. Knowledge-based visual QA established that many visual questions require facts the image does not contain (Marino et al., 2019; Schwenk et al., 2022; Chang et al., 2022), and multimodal search benchmarks now evaluate agents that retrieve them (Jiang et al., 2024; Li et al., 2024; Tao et al., 2026; Zeng et al., 2026). All of them score the final answer, and recency is the most common proxy for search difficulty (Vu et al., 2024; Kasai et al., 2023). LiveVQA filters by source-context answerability and model-based tests (Fu et al., 2025), MMSearch-Plus removes items that tested models solve without search (Tao et al., 2026), and VDR-Bench screens textual and prior-based shortcuts (Zeng et al., 2026). Our distinction is not a stricter gate but a retained one: we make the joint admission outcome an explicit per-item predicate over repeated trials,

Table 1: **Comparison with existing benchmarks.** ✓ denotes the stated criterion; ✓* a restricted or qualified check; × that the mechanism is not established in the inspected paper. Search and oracle columns are panel- and protocol-relative admission checks; Support × correctness requires joint classification of observed support and answer correctness for correct and incorrect runs.

Benchmark	Visual input	Live renewal	Search necessity	Oracle answerability	Support × correctness	Human review
MMSearch (Jiang et al., 2024)	✓	✓*	✓*	×	×	✓
Dyn-VQA (Li et al., 2024)	✓	✓	×	×	×	✓
LiveVQA (Fu et al., 2025)	✓	✓	✓*	✓*	×	✓
MMSearch-Plus (Tao et al., 2026)	✓	✓	✓*	×	×	✓
LiveNewsBench (Zhang et al., 2026)	×	✓	×	✓*	×	✓
VDR-Bench (Zeng et al., 2026)	✓	×	×	×	×	✓
LIVESEARCHVQA (ours)	✓	✓	✓	✓	✓	✓

keep the trials and the reference evidence in the item record, and reuse that same evidence to diagnose search outcomes at run level.

Diagnosing what an outcome reveals. KILT made answer provenance an evaluation target (Petroni et al., 2021), RAGChecker reports retrieval and generation diagnostics separately (Ru et al., 2024), and SealQA shows that retrieved context can remain difficult to use (Pham et al., 2026); related work isolates distraction, evidence position, and knowledge conflict (Shi et al., 2023; Liu et al., 2024; Yoran et al., 2024b; Wang et al., 2024). Intermediate-entity recall and tool-use error analyses further characterize parts of the search process (Zeng et al., 2026; Tao et al., 2026). What they do not establish, per run, is whether sufficient support for the validated reference fact was present in the agent-visible trace. LIVESEARCHVQA takes that reference from question construction and applies it to the content each agent actually rendered, so that observed support and answer correctness are classified jointly.

Renewable evaluation. Periodically refreshed collections limit exposure to prior training data (White et al., 2025; Jain et al., 2025; Zhang et al., 2026), and iterative construction sustains evaluation as models improve (Kiela et al., 2021); contamination studies caution that exposure is specific to a model and a benchmark rather than implied by a publication date (Golchin & Surdeanu, 2024; Sainz et al., 2023). Renewal alone does not preserve interpretability: a new question is informative only if it still passes the checks that made earlier ones informative. We therefore repeat the same admission predicate on every dated build and retain enough of each build's construction record that later support diagnosis still targets the fact validated at that time. Table 1 summarizes these distinctions, and Appendix H extends the comparison.

3 PROBLEM FORMULATION AND BENCHMARK DESIGN

3.1 TASK, SCOPE, AND EVALUATION CONDITIONS

We study time-anchored visual search questions in which an image identifies the search target and external sources provide the requested fact. Each question names an event or explicit reference date, so its answer is fixed at construction. Our goal is to evaluate two things that a final answer conflates: whether the question tests an information gap, and whether the agent obtained sufficient support before answering. For item i under condition z, an agent produces

$$(\hat{a}_i^z, H_i^z) \sim \pi(\cdot \mid I_i, q_i, E_i^z;\ \mathcal{T}^z), \qquad z \in \{\mathrm{CB}, \mathrm{WS}, \mathrm{OR}\}, \tag{1}$$

where $\hat{a}_i^z$ is the answer, H_i^z the trace of tool calls and tool-returned content rendered to the agent, and E_i^z the evidence supplied in advance, and $\mathcal{T}^z$ the available tools. Closed book (CB) provides

neither evidence nor tools; with search (WS) provides tools but no prior evidence; oracle evidence (OR) provides the packet E_i^* and no tools, so its trace is empty.

The benchmark additionally stores a reference answer a_i^*, the supporting span e_i^*, and source, timing, and version records; these serve construction and scoring, not the agent. The packet E_i^* pairs e_i^* with its source title, so an oracle answer is conditioned on the packet rather than the bare span.

OR is a reference condition rather than an upper bound on WS, since a web result may be clearer than the recorded packet. Admission uses CB and OR screening trials, while all reported scores come from fresh runs(Appendix A).

3.2 Question validity: establishing an information gap

We first check that a candidate is well formed: its visual referent and reference time are identifiable, the recorded source supports the reference answer, and that answer is not directly readable from the image. Structural checks on source freshness, image usability, span occurrence, answer–span consistency, and identifier uniqueness, together with the visual-relevance judgments, produce the validity indicator V_i (Appendix A). Validity in this sense does not show that answering requires search, which holds only relative to a set of models and what they are given. Let $C_{imr}^z \in \{0,1\}$ denote the correctness of admission model m on item i in trial r under $z \in \{\mathrm{CB}, \mathrm{OR}\}$, for a fixed admission panel $\mathcal{M}$ and R trials per condition. Admission requires

$$B_i = \bigwedge_{m\in\mathcal{M},\, r\le R} \big[C_{imr}^{\mathrm{CB}} = 0\big], \quad O_i = \bigwedge_{m\in\mathcal{M},\, r\le R} \big[C_{imr}^{\mathrm{OR}} = 1\big], \quad \mathrm{Accept}_i = V_i \wedge B_i \wedge O_i, \tag{2}$$

where B_i denotes repeated closed-book failure and O_i denotes repeated supplied-evidence success. Requiring both separates an information gap from mere difficulty: B_i alone does not distinguish a recoverable gap from a question that remains unanswerable even with the supporting passage, while O_i alone would admit questions every model already answers.

Abstention counts as an incorrect closed-book answer; the certificate is panel-relative rather than a claim of universal ignorance; and V_i establishes visual *relevance* rather than image *necessity* (Appendix A.2; tested in Sections 4.2 and 5.1). Admission establishes a question-level information gap; it does not establish what evidence a search agent will actually obtain.

3.3 Outcome diagnosis: evidence access and use

Diagnosis applies to a single with-search run, not to a question: the same item reaches different agents through different queries, tool profiles, and truncation boundaries. For a given run, let $C = 1$ indicate a correct answer and $G = 1$ indicate that sufficient support for the reference fact appeared in content visible to the agent before it answered. Validated equivalent sources count, since an agent that finds the same fact elsewhere has not failed to obtain it; a bare URL, an unopened page, and a passage removed by truncation do not, since none entered the agent-visible context. G records support in agent-visible content, not that a relevant document existed in a result list or that the agent found material consistent with a wrong answer. Crossing the labels gives four outcomes. Errors without observed support direct attention to query formation, page access, and truncation; errors despite support direct it to referent resolution, competing facts, and answer selection. The cells locate an error relative to the evidence but do not identify its cause. Correct answers without observed support are equally informative and equally underdetermined: prior knowledge, a lucky guess, or an equivalent source the matcher missed. On runs whose support state is resolved, write $\mathrm{Cov} = \Pr(G = 1)$ for coverage, $U_1 = \Pr(C = 1 \mid G = 1)$, and $U_0 = \Pr(C = 1 \mid G = 0)$; these reconstruct answer accuracy exactly:

$$\mathrm{Acc}_{\mathrm{WS}}^{\mathrm{resolved}} = \mathrm{Cov}\, U_1 + (1 - \mathrm{Cov})\, U_0. \tag{3}$$

Ambiguous traces keep an unresolved label with its own denominator rather than being recoded as $G = 0$ (Appendix C). U_1 is correctness conditional on observed support, not a causal claim about evidence use; Section 5.4 tests the implied hypotheses.

3.4 Live renewal: new questions, fixed references

Each build admits new time-anchored questions under Eq. 2 and retains the reference records required for run-level diagnosis (Section 3.3). A new event or reference time creates a new item; released

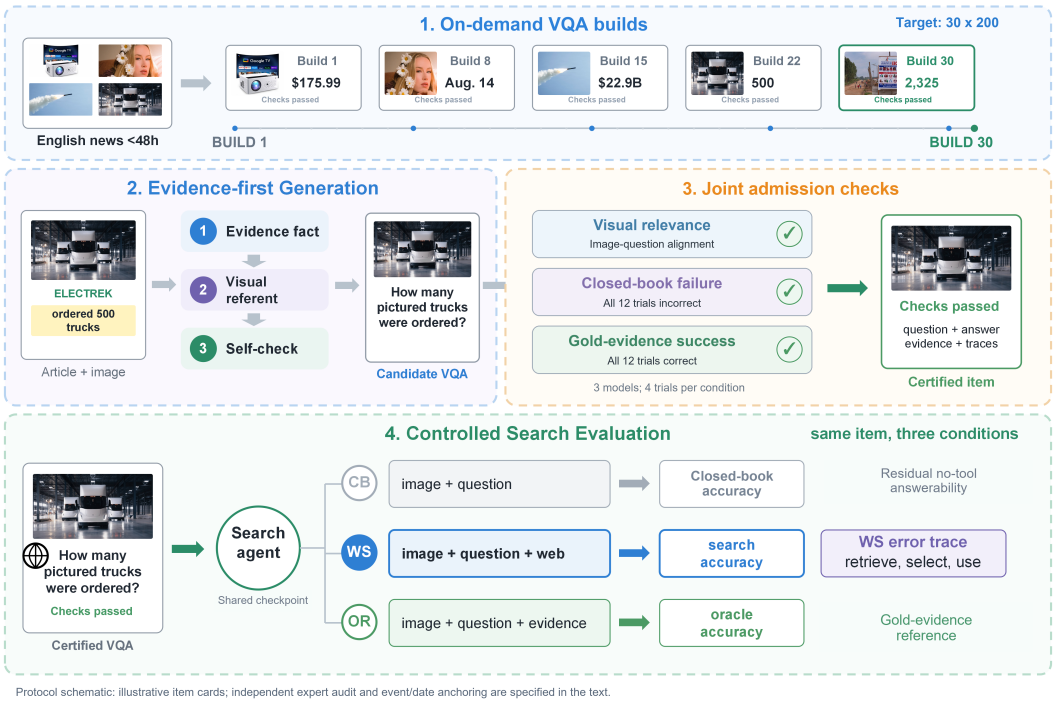

Figure 2: **One evidence record connects construction and evaluation.** Recent source facts ground candidate visual questions; joint checks retain the evidence and trial records; paired CB/WS/OR evaluation then connects accuracy to evidence access and use.

versions preserve their questions, answers, reference times, supporting evidence, and evaluation records. Confirmed defects create superseding versions rather than in-place edits.

Comparisons fix item versions, evaluation windows, and tool profiles. Builds are not calibrated to a common difficulty scale, so cross-build score differences alone do not establish model progress. Timestamp definitions and the publication-to-build freshness constraint are detailed in Appendix F.

4 Benchmark Construction and Evaluation

One versioned evidence record connects question admission to run-level support diagnosis (Figure 2).

4.1 Evidence-first construction and joint admission

Generation. The generator selects an event-specific fact and its exact supporting span from a recent article, then writes a question linking that fact to a visual subject in an accompanying image. The retained span makes source support checkable before admission.

Joint admission. Visual screening uses image descriptions: two separately prompted judgments must agree on image–source alignment and question–image grounding at or above an identifiable referent. These checks establish visual relevance; original-image necessity is assessed separately.

Candidates then undergo Eq. 2 in fresh contexts. Three endpoints from two model families each contribute four trials per condition, yielding twelve CB and twelve OR responses per admitted item. CB respondents see only the image and question; OR additionally supplies E_i^*. Reference answers, screening histories, and other respondents' outputs are withheld. The same versioned equivalence rule scores admission and evaluation, preserving a common definition of correctness. Retained records include supporting spans, source archives, and screening responses, which are not supplied in WS runs. Panel identities, generator-family overlap, prompts, and short-circuit execution are documented in Appendices A and B.

4.2 LIVE RENEWAL AND RELEASE VALIDATION

Builds. Each build draws candidates from sources satisfying $0 \leq t_{\text{build}} - t_{\text{pub}} \leq 48$ hours and applies admission unchanged; composition caps never relax a threshold to reach a target count. Across thirty dated builds, 16,401 candidates yielded 7,861 admitted and 6,000 released items over five topic domains and five answer types, with admitted counts exceeding the release cap in every build (Appendix F).

Release validation. A domain-stratified human audit of 600 released items, held-out-model CB/OR evaluation, and visual controls assess properties beyond admission (Section 5.1). Reviewers lock question-only judgments before viewing the image (Appendix D). Released references and supersession records are preserved under Section 3.4.

4.3 PAIRED EVALUATION AND EVIDENCE ANNOTATION

Paired runs and scoring. Fresh CB, WS, and OR runs use the same item versions, model revisions, answer-format instructions, and scoring rule. Common-scaffold systems share tools and budgets; native-tool systems form a separate reporting group. Typed equivalence rules preserve the requested relation, unit, sign, scale, and date; residual semantic cases reach a separate grader (Appendix C). Correctness and evidence sufficiency are assessed separately.

Support labels. A frozen matcher inspects content rendered before the final answer and checks support for the reference entity, relation, event or date, and value. Every positive label must cite a passage in the original visible trace: the retained reference defines sufficient evidence but does not establish that the agent saw it. We denote the automatic label by $\widehat{G}$. If trace visibility or evidence ambiguity prevents a support decision, the label remains unresolved rather than being assigned $G = 0$ (Appendix C).

Controlled comparisons. Each hypothesis the four cells suggest is tested by holding the question fixed and changing one property of the evidence: matched irrelevant or conflicting passages added to the oracle packet, retrieval depth and answer-blind reranking on a fixed snapshot, and visual controls that remove the image or replace it with text naming the referent. Every comparison retains corrected and newly wrong responses separately, and retrieval comparisons report coverage alongside accuracy, because removing context can remove support as well as noise.

5 EXPERIMENTS

Four questions follow from the design: validity (Section 5.1), label reliability (Section 5.2), what the labels add to a final-answer score (Section 5.3), and whether those directions survive a controlled test (Section 5.4). All results use the twelve configurations in Table 5 on the same 6,000 items. Every interval uses 2,000 paired cluster-bootstrap resamples over 30 builds (Efron & Tibshirani, 1993); both sides of every contrast rest on the same resampled questions (Appendix D.2).

5.1 BENCHMARK VALIDITY UNDER RENEWAL

The released questions meet the task definition. Three calibrated reviewers independently assessed a domain-stratified sample of 600 released items against six criteria, blinded to admission verdicts, with text-only judgments locked before the image was revealed. Of the 600, 599 pass all six criteria and one remains unresolved, with no confirmed defect. A sample of this size detects a 1% defect rate with probability above 95% even when a sampled defect is recognized only 80% of the time (Appendix D).

Held-out answerability. Six systems outside the admission families average 9.7% CB and 93.6% OR accuracy, compared with 10.1% and 93.1% for six Qwen-family systems (Table 9). These results show low closed-book and high supplied-evidence answerability beyond the admission panel. They concern marginal accuracy rather than transfer of the repeated per-item admission certificate. External-system means remain at 8.6–10.8% CB and 92.0–95.0% OR across the observed builds (Appendix F).

Table 2: **Answer accuracy and evidence diagnostics on LIVESEARCHVQA.** All systems are evaluated on the same 6,000 items; values are percentages. Coverage is the fraction of WS runs with sufficient evidence visible before the final answer; U_1 and U_0 are correctness with and without observed support. Intervals and per-build results are in Appendix D.2.

	Answer accuracy			Evidence diagnostics		
System	CB	WS	OR	Coverage	U_1	U_0
Frontier reference: native tool profile						
claude-opus-5	9.5	61.9	94.4	74.0	80.9	7.6
gpt-5.5	10.0	60.6	95.0	70.0	82.6	9.3
gemini-3.1-pro	8.4	57.6	94.1	74.6	75.1	6.3
Common ReAct scaffold: shared tool interface and budget						
qwen3.7-plus	10.0	58.1	93.9	66.9	81.9	9.7
GLM-5V-turbo	8.9	53.8	93.5	71.9	71.9	7.5
qwen3.7-flash	11.2	53.8	94.2	63.1	78.9	10.9
Qwen3-VL-32B	8.7	50.7	93.5	73.1	66.5	7.9
qwen3-vl-235b	10.9	51.0	92.7	66.1	71.9	10.2
Llama-4-Maverick	9.5	46.7	92.7	67.1	65.5	8.3
Qwen3-VL-8B	10.4	47.1	92.5	60.4	70.8	10.8
Qwen3-VL-30B-A3B	9.6	43.6	91.8	67.8	59.9	9.4
Llama-4-Scout	11.7	44.5	92.0	62.3	64.1	12.3

The image determines what the agent searches for. Four configurations received the same 6,000 questions under three matched inputs, with a separate target-identification probe run without search (Appendix E). Removing the image reduces target identification from 96–98% to 15–21% and with-search accuracy from 46.7–61.9% to 16.0–18.0%, retaining only 25.0% of the originally correct answers. Naming the referent in text, without revealing the requested fact, moves the other way: with-search accuracy rises to 69–77% and 95.0% of original successes are retained, while closed-book accuracy rises only from 8.4–10.0% to 10.5–12.7% — 16.3 points of with-search gain against 2.4 points of closed-book gain. Residual closed-book successes in Table 2 bound the search requirement for the evaluated configurations; the item set was frozen before any evaluated system was run.

5.2 RELIABILITY OF EVIDENCE DIAGNOSIS

The analyses that follow read the support label G, so its reliability is a precondition rather than a side result. We evaluate it where a wrong label would be most consequential: an incorrect answer, where the difference between an error without observed support and an error despite it changes what a developer would do next. On a fixed cohort of 600 incorrect with-search runs, evenly split between independently adjudicated supported and unsupported errors, two diagnostic conditions receive the same question, image, reference answer, final response, and complete pre-answer trace; only the record-aware condition additionally receives the retained supporting passage and its source metadata. That record states what evidence would have sufficed and is not treated as evidence the agent saw, so a positive decision must still cite support inside the original visible trace. Balanced accuracy rises from 80.0% to 93.0%, a 65% reduction in balanced diagnostic error, with class recalls improving from 78% and 82% to 92% and 94% (Appendix G.2). The search answers are identical in both conditions; what changes is how reliably each failure can be attributed to an evidence state.

The result certifies support-state classification among failures, not the full four-cell assignment; support labels on correct answers are sampled separately, and ambiguous traces keep an unresolved label.

5.3 AGENT PERFORMANCE AND EVIDENCE PROFILES

Table 2 reports WS accuracy of 43.6–61.9%, compared with 91.8–95.0% under OR. The corresponding systems answer substantially more questions when recorded evidence is supplied directly. Joint

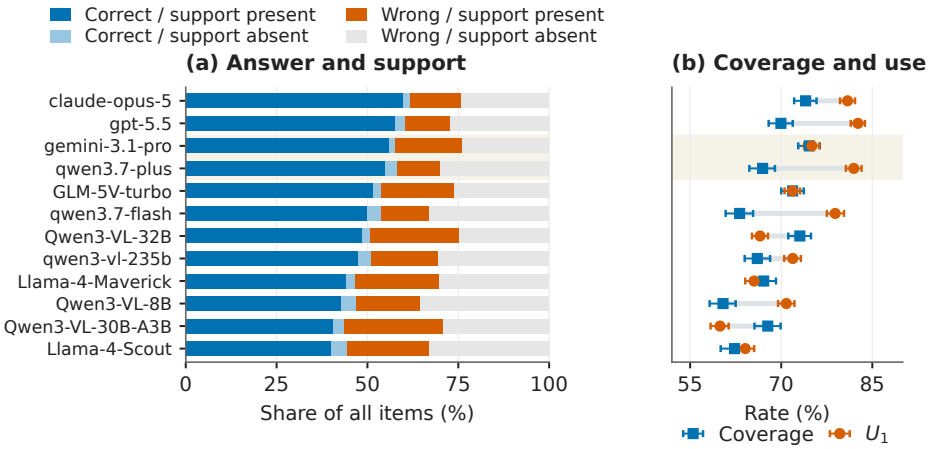

Figure 3: **Similar accuracy can conceal different evidence profiles.** (a) The four evidence–correctness outcomes partition all 6,000 runs for each system. (b) Coverage uses all runs, whereas supported accuracy uses only runs with sufficient evidence. The highlighted gemini-3.1-pro/qwen3.7-plus pair reverses order across the two measures; the same reversal holds for GLM-5V-turbo and qwen3.7-flash under a single declared tool profile (Section 5.3).

evidence–correctness counts distinguish outcomes with and without observed support (Figure 3; Appendix D.2).

Systems with statistically indistinguishable scores make the separation concrete. Within the common-scaffold group, GLM-5V-turbo and qwen3.7-flash differ by 0.07 percentage points in with-search accuracy, an interval that includes zero. Their evidence profiles are nevertheless clearly ordered and in opposite directions: sufficient evidence reaches GLM-5V-turbo on 71.9% of items against 63.1% for qwen3.7-flash, while correctness among supported runs is 71.9% against 78.9%. The gemini-3.1-pro/qwen3.7-plus pair shows the same pattern at a 0.45-point difference (Figure 3). The same score is reached from opposite directions — broader observed support against higher correctness given support — and records neither component.

Correct answers without observed support must stay outside this comparison. They account for 6.3% to 12.3% of unsupported runs, and folding them into supported accuracy by reporting $\mathrm{Acc}_{\mathrm{WS}}/\mathrm{Cov}$ inflates it by up to 7.4 percentage points (Llama-4-Scout).[1] These profiles imply different priorities: errors without observed support point to retrieval and evidence access, errors despite support to evidence interpretation and answer selection. The next section tests both.

5.4 Controlled retrieval and context interventions

Four configurations undergo paired context and retrieval interventions on the same 6,000 items (Table 12). Retrieval comparisons use a fixed snapshot; reranking has no access to gold answers (Appendix G).

Added context. Appending five irrelevant passages to the oracle packet changes accuracy by only -0.2 to -1.3 pp, whereas replacing them with five conflicting passages of matched length, source cues, and gold position costs a further -6.9 to -14.0 pp (Figure 4a). Longer context is not the operative factor: the reference fact remains visible in both conditions by construction.

Retrieval depth. Increasing depth from three to twenty results raises coverage for every configuration, yet with-search accuracy rises by 11.0 pp for claude-opus-5 and 11.8 pp for qwen3.7-plus, but falls by 1.4 pp for Qwen3-VL-32B (Figure 4c,d). A gain in Cov can be cancelled by a loss in U_1.

[1] For gemini-3.1-pro, 3,360 of 4,477 supported runs are correct and a further 96 are correct without observed support, so $\mathrm{Acc}_{\mathrm{WS}} = (3360 + 96)/6000 = 57.6\%$ and $U_1 = 3360/4477 = 75.1\%$, whereas $\mathrm{Acc}_{\mathrm{WS}}/\mathrm{Cov} = 3456/4477 = 77.2\%$ (Table 10).

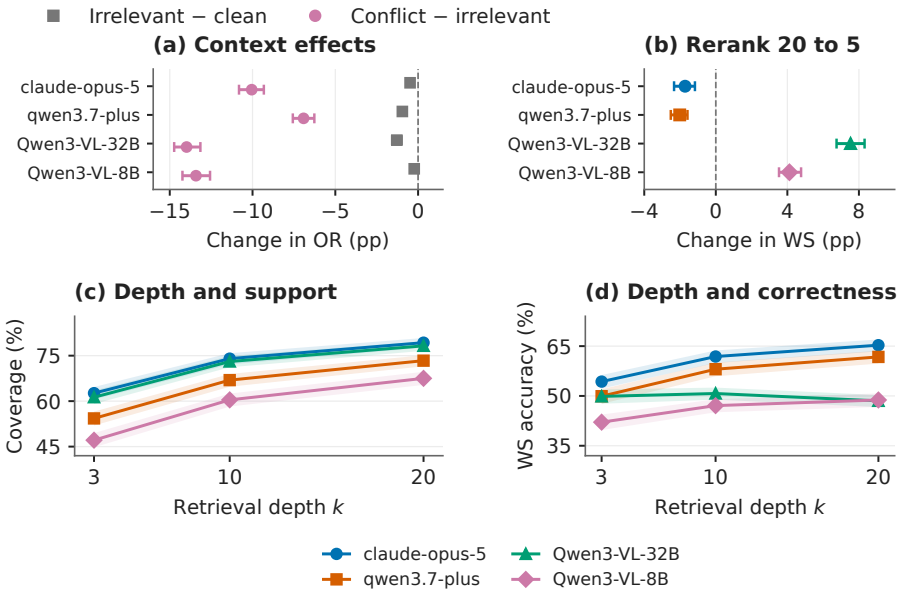

Figure 4: **Access and correctness respond differently to evidence interventions.** (a) Added irrelevant context and the additional conflict effect. (b) Reranking from 20 context units to five. (c,d) Retrieval depth changes support and correctness separately.

Reranking. Reducing the same twenty-result pool to five with an answer-blind reranker reverses sign: -1.7 pp for claude-opus-5 and -2.0 pp for qwen3.7-plus, against $+7.5$ pp for Qwen3-VL-32B and $+4.1$ pp for Qwen3-VL-8B. Net figures understate churn: claude-opus-5's -1.7 pp combines 80 corrected errors with 183 newly introduced ones. Coverage is therefore reported alongside accuracy for every retrieval condition.

Access and correctness thus respond to different manipulations, and a system's position on one axis does not predict its response on the other. Appendix G.1 uses the development-set diagnosis to choose between these interventions before test outcomes are revealed.

6 Limitations and Conclusion

LiveSearchVQA connects two questions that end-to-end accuracy leaves open: why a question should test live search, and what evidence was available when an agent answered it. Evidence-first construction and joint admission make the first decidable; paired conditions and trace accounting separate the second from answer correctness. Systems with statistically indistinguishable scores reach them through opposite combinations of evidence access and supported correctness, and interventions that increase access change accuracy in both directions. Because each dated build must establish validity under the same rule and retain the same references, later cohorts remain comparable to earlier ones.

Three limitations bound these claims. The collection covers recent English-language, event-specific questions. Admission is panel- and protocol-relative, and live retrieval means comparisons remain tied to recorded item versions and tool profiles. Support judgments involve interpretation when evidence is distributed or only partly exposed; we report these cases separately rather than resolving them by convention.

AI USE STATEMENT

We used generative AI tools to organize the manuscript structure, to revise and polish the writing, and to discuss experimental-design options during development (for example, which analyses to prioritize and how to present them). All AI-assisted text and suggestions were reviewed and verified by the authors. Model endpoints that appear in construction, admission, grading, and evidence diagnosis are part of the method and are specified in the paper. We take responsibility for the final content of this work.

REFERENCES

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.11511.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022. URL https://arxiv.org/abs/2212.08073.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. WebQA: Multihop and multimodal QA. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16495–16504, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Chang_WebQA_Multihop_and_Multimodal_QA_CVPR_2022_paper.html.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 14948–14968, 2023. URL https://aclanthology.org/2023.emnlp-main.925/.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2Web: Towards a generalist agent for the web. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.52202/075280-1220. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/5950bf290a1570ea401bf98882128160-Abstract-Datasets_and_Benchmarks.html.

Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1923, 1998. doi: 10.1162/089976698300017197. URL https://doi.org/10.1162/089976698300017197.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993. doi: 10.1007/978-1-4899-4541-9. URL https://doi.org/10.1007/978-1-4899-4541-9.

Mingyang Fu, Yuyang Peng, Dongping Chen, Zetong Zhou, Benlin Liu, Yao Wan, Zhou Zhao, Philip S. Yu, and Ranjay Krishna. Seeking and updating with live visual knowledge. *arXiv preprint arXiv:2504.05288*, 2025. URL https://arxiv.org/abs/2504.05288v2.

Shahriar Golchin and Mihai Surdeanu. Time travel in LLMs: Tracing data contamination in large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/pdf?id=2Rwq6c3tvr.

Hongliang He, Wenlin Yao, Kaixin Ma, Wenhao Yu, Yong Dai, Hongming Zhang, Zhenzhong Lan, and Dong Yu. WebVoyager: Building an end-to-end web agent with large multimodal models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6864–6890, 2024. doi: 10.18653/v1/2024.acl-long.371. URL https://aclanthology.org/2024.acl-long.371/.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida I. Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/94074dd5a072d28ff75a76dabed43767-Abstract-Conference.html.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, Peng Gao, Yu Liu, Chunyuan Li, and Hongsheng Li. MMSearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*, 2024. URL https://arxiv.org/abs/2409.12959v2.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. RealTime QA: What's the answer right now? In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9941624ef7f867a502732b5154d30cb7-Abstract-Datasets_and_Benchmarks.html.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4110–4124, 2021. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324/.

Jing Yu Koh, Robert Lo, Lawrence Jang, Vikram Duvvur, Ming Chong Lim, Po-Yu Huang, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried. VisualWebArena: Evaluating multimodal agents on realistic visually grounded web tasks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 881–905, 2024. URL https://aclanthology.org/2024.acl-long.50.pdf.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. Vlind-bench: Measuring language priors in large vision-language models. *Findings of the Association for Computational Linguistics: NAACL*, 2025. URL https://arxiv.org/abs/2406.08702.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose G. Moreno, and Jesús Lovón Melgarejo. ViQuAE, a dataset for knowledge-based visual question answering about named entities. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3108–3120, 2022. URL https://github.com/PaulLerner/ViQuAE.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Shilong Li, Xingyuan Bu, Wenjie Wang, Jiaheng Liu, Jun Dong, Haoyang He, Hao Lu, Haozhe Zhang, Chenchen Jing, Zhen Li, Chuanhao Li, Jiayi Tian, Chenchen Zhang, Tianhao Peng, Yancheng He, Jihao Gu, Yuanxing Zhang, Jian Yang, Ge Zhang, Wenhao Huang, Wangchunshu Zhou, Zhaoxiang Zhang, Ruizhe Ding, and Shilei Wen. MM-BrowseComp: A comprehensive benchmark for multimodal browsing agents, 2025. URL https://arxiv.org/abs/2508.13186v1.

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Philip S. Yu, Fei Huang, and Jingren Zhou. Benchmarking multimodal retrieval augmented generation with dynamic VQA dataset and self-adaptive planning agent. *arXiv preprint arXiv:2411.02937*, 2024. URL https://arxiv.org/abs/2411.02937.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3195–3204, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Marino_OK-VQA_A_Visual_Question_Answering_Benchmark_Requiring_External_Knowledge_CVPR_2019_paper.html.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3113–3124, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Mensink_Encyclopedic_VQA_Visual_Questions_About_Detailed_Properties_of_Fine-Grained_Categories_ICCV_2023_paper.pdf.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. *arXiv preprint arXiv:2311.12983*, 2023. URL https://arxiv.org/abs/2311.12983.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. URL https://arxiv.org/abs/2112.09332.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2103.14749.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*, 2024. URL https://arxiv.org/abs/2404.13076.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. KILT: A benchmark for knowledge intensive language tasks. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 2523–2544, 2021. doi: 10.18653/v1/2021.naacl-main.200. URL https://aclanthology.org/2021.naacl-main.200/.

Thinh Pham, Nguyen Nguyen, Pratibha Zunjare, Weiyuan Chen, Yu-Min Tseng, and Tu Vu. SealQA: Raising the bar for reasoning in search-augmented language models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2506.01062v4.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of EMNLP*, pp. 5687–5711, 2023. doi: 10.18653/v1/2023.findings-emnlp.378. URL https://aclanthology.org/2023.findings-emnlp.378/.

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Stéphane Clinchant, and Vassilina Nikoulina. BERGEN: A benchmarking library for retrieval-augmented generation. In *Findings of EMNLP*, pp. 7640–7663, 2024. doi: 10.18653/v1/2024.findings-emnlp.449. URL https://aclanthology.org/2024.findings-emnlp.449/.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Jiayang Cheng, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.52202/079017-0692. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/27245589131d17368cccdfa990cbf16e-Abstract.html.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of EMNLP*, pp. 10776–10787, 2023. URL https://aclanthology.org/2023.findings-emnlp.722/.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. doi: 10.52202/075280-2997. URL https://proceedings.neurips.cc/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision (ECCV)*, pp. 146–162, 2022. URL https://link.springer.com/chapter/10.1007/978-3-031-20074-8_9.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning (ICML)*, pp. 31210–31227, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. MultiModalQA: Complex question answering over text, tables and images. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2104.06039.

Xijia Tao, Yihua Teng, Xinxing Su, Xinyu Fu, Jihao Wu, Chaofan Tao, Ziru Liu, Haoli Bai, Rui Liu, and Lingpeng Kong. MMSearch-Plus: Benchmarking provenance-aware search for multimodal browsing agents. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2508.21475v3.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720, 2024. URL https://aclanthology.org/2024.findings-acl.813/.

Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/pdf?id=ptvV5HGTNN.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. BrowseComp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025. URL https://arxiv.org/abs/2504.12516.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/e4a46394ba5378b3f9a186a5b4c650d1-Abstract-Conference.html.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. WebWalker: Benchmarking LLMs in web traversal. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10290–10305, 2025. doi: 10.18653/v1/2025.acl-long.508. URL https://aclanthology.org/2025.acl-long.508/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.03629.

Ori Yoran, Samuel Joseph Amouyal, Chaitanya Malaviya, Ben Bogin, Ofir Press, and Jonathan Berant. AssistantBench: Can web agents solve realistic and time-consuming tasks? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8938–8968, 2024a. doi: 10.18653/v1/2024.emnlp-main.505. URL https://aclanthology.org/2024.emnlp-main.505/.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations (ICLR)*, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8011b23e1dc3f57e1b6211ccad498919-Abstract-Conference.html.

Yu Zeng, Wenxuan Huang, Zhen Fang, Shuang Chen, Yufan Shen, Yishuo Cai, Xiaoman Wang, Zhenfei Yin, Lin Chen, Zehui Chen, Shiting Huang, Yiming Zhao, Xu Tang, Yao Hu, Philip Torr, Wanli Ouyang, and Shaosheng Cao. Vision-deepresearch benchmark: Rethinking visual and textual search for multimodal large language models, 2026. URL https://arxiv.org/abs/2602.02185.

Yunfan Zhang, Kathleen McKeown, and Smaranda Muresan. LiveNewsBench: Evaluating LLM web search capabilities with freshly curated news, 2026. URL https://arxiv.org/abs/2602.13543.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4410c0711e9154a7a2d26f9b3816d1ef-Abstract-Conference.html.

A ITEM RECORDS AND CONSTRUCTION PROFILES

The construction record links each decision to the item and evidence that produced it. It contains the item version, image or authorized reference, source URL and hash, archived text, publication and build timestamps, event identifier, question, answer type, canonical answer, accepted aliases, and supporting offsets. Verification fields retain the generator and rule versions, rejection stage, visual judgments, panel endpoints, prompts, decoding settings, responses, grader version, and trial verdicts.

Deterministic checks. Checks cover source freshness, image usability, exact span occurrence, answer–span consistency, and unique item IDs. Numeric matching preserves the requested relation: a price, percentage, date, and quantity may appear in the same article. Source titles or surrounding sentences form part of OR only when included in the declared evidence packet.

Original visual relevance rubric. Image–source alignment uses five ordered scores: 0, contradictory; 1, unrelated or generic; 2, plausible but ambiguous; 3, correct entity with unresolved event association; and 4, verified correspondence at the level required by the question. The threshold is 4. Question–image grounding uses 0, irrelevant image; 1, decorative association; 2, identifiable referent; 3, disambiguation among plausible referents; and 4, fine-grained visual cues needed for resolution. The original threshold is 2. Two separately prompted referent judgments must agree. These judgments use the image description; these checks establish visual relevance. The stronger profile below requires the original pixels and tests necessity directly.

Generation profile. Candidate facts, supporting spans, and questions were produced by `doubao-seed-2.0-pro` under a fixed generator prompt and version within each build. The generator reads the source article; it never sees admission outcomes, and a proposal carries no admission privilege.

Executed admission panel. Admission used three endpoints, `doubao-seed-pro-2.1`, `qwen3.8-max`, and `qwen3.8-27B`, spanning two model families. Each endpoint contributed four fresh-context trials per condition, so every admitted item retains $3 \times 4 = 12$ closed-book and twelve oracle-evidence responses. Table 3 lists the endpoints and their decoding settings: closed-book trials keep each endpoint's default sampling temperature so that repeated trials expose the variation a user would encounter, whereas oracle-evidence trials use a common low temperature so that supplied-evidence success reflects the evidence packet rather than sampling noise. Respondents never see the reference answer, admission histories, generator identities, or one another's responses; closed-book trials additionally exclude the article title and any source snippet, and the oracle condition adds only the declared evidence packet E^*. The profile permits UNKNOWN, which the scorer marks incorrect, so the closed-book check certifies the absence of an expressed correct answer rather than the absence of the fact from a model's parameters.

Table 3: **Executed admission panel and decoding settings.** Each endpoint contributes four fresh-context trials per condition, so every admitted item retains twelve closed-book and twelve oracle-evidence responses.

Endpoint	Family	CB temp.	OR temp.	top-p	Trials / condition
doubao-seed-pro-2.1	Doubao	0.8	0.2	0.80	4
qwen3.8-max	Qwen	0.7	0.2	0.80	4
qwen3.8-27B	Qwen	0.7	0.2	0.95	4
Retained per admitted item		12 CB responses, 12 OR responses			

Evidence that the certificate transfers beyond the panel lineage. The admitted questions behave as intended for families unrelated to the panel. Across the six evaluated systems from Anthropic, OpenAI, Google, Zhipu, and Meta in Table 9, closed-book accuracy stays between 8.4% and 11.7% and oracle-evidence accuracy between 92.0% and 95.0%. A selection effect specific to the panel would predict the opposite pattern within the roster, namely lower closed-book accuracy for systems sharing the panel's lineage. The six Qwen-family systems instead average 10.1% closed-book and

93.1% oracle-evidence accuracy, against 9.7% and 93.6% for the six unrelated systems. Independent human review (Appendix D) assesses the same items without any panel involvement.

Answer equivalence. Scoring normalizes Unicode, whitespace, currency symbols, and declared unit aliases, then parses the answer type. It preserves currency, scale, sign, dates, fiscal periods, percentage versus percentage points, and current versus list price. Exact quantities require canonical equality; rounding tolerance follows the question's stated precision. Incompatible alternatives make an answer incorrect even if one matches the gold. Remaining semantic cases use a versioned rubric with human escalation and retained explanations.

Grader identity and separation. The equivalence rule $\mathrm{Eq}(\hat{a}, a^*; q, e^*)$ supplies the single definition of correctness used both to admit questions in Eq.(2) and to score every reported CB, WS, and OR run, so the identity of the grader is part of the protocol rather than an implementation detail. Grading proceeds in two stages. The deterministic typed rules above resolve most comparisons, and only residual semantic cases reach a model grader; its influence is therefore bounded by an escalation rate that we report alongside the accuracy tables. We use `grok-4.6` in that role because no xAI model has any other role in this work: it does not generate candidates, does not sit on the admission panel, and does not appear among the evaluated systems (Table 4 and Table 5) or the configurations used for the visual and evidence controls. A grader drawn from an evaluated family would allow self-preference to act on reported scores rather than only on item selection, which is why this role is separated even though the construction panel itself is not. The grader is frozen within a release: a fixed model version, rubric version, prompt, decoding configuration, and output schema are applied in a fresh context for each decision. It receives the question, the reference answer with its accepted aliases, the candidate answer, and, where the rubric requires disambiguation, the supporting span; it does not receive the respondent's identity, the condition label, the admission history, or any other response to the same item. Decisions marked uncertain, together with a fixed random sample of the remainder, receive human adjudication with retained explanations and reported agreement. The trace matcher that produces $\hat{G}$ and the diagnostic judge in Appendix G.2 are separate frozen components governed by the same separation rule, and their records are kept apart from grading records.

Selection and maintenance. Near-duplicate images, paraphrases, and syndicated stories are grouped by event. Release selection uses fixed composition caps and priorities without relaxing validity thresholds to reach a target count. A later fact creates a new time-anchored question; a correction creates a superseding version. Panel agreement and admission rate describe selection, while independent review assesses whether the selected items meet the task definition.

A.1 Family separation and strict visual admission

Roles and information boundaries. The strengthened profile requires five model families: one generator, three admission families, and one external family. Generation uses a fixed source pool with proposal budgets stratified by date, domain, and answer type. The generator family is excluded from admission; the three admission members must represent three distinct families, rather than three sizes or endpoints from one family. Structured answers use deterministic equivalence rules; remaining semantic cases receive independent human grading and adjudication, without adding another model family. One external family has no role in generation, prompt selection, threshold calibration, or rejection feedback. A second external family and additional generator families are optional extensions, subject to the same separation rule. Respondents never see reference answers, admission histories, generator identities, or each other's responses. Closed-book trials also exclude article titles and source snippets. Four repetitions per condition remain useful for stochastic variation but are not counted as additional families. Family membership, available training-lineage information, version identifiers, and any unavoidable role overlap are recorded explicitly.

Calibration and panel ablations. Family separation addresses a potential source of correlated judgments, including self-preference effects documented in LLM evaluation (Panickssery et al., 2024); it does not guarantee independence. Before the test period, a separate event-disjoint, human-labelled set contains valid questions and targeted invalid controls: ambiguous referents, text-recoverable targets, unsupported answers, outdated facts, and pixel leakage. Each verifier is evaluated against these labels. False-acceptance and false-rejection rates are reported separately with their denominators.

Pairwise shared errors are measured within validity, domain, answer-type, and human-difficulty strata; residual error correlations assess redundancy beyond differences in case mix. Agreement alone does not establish correctness. Panel selection and thresholds are fixed using this calibration set, without optimizing leaderboard scores or transfer on the external families.

A common cached candidate pool supports three comparisons: one verifier, three endpoints from one family, and three distinct-family verifiers. The comparisons retain the same prompts, repetitions, grader, and candidate budgets, and report admission counts, independently reviewed validity, false rejection, composition, and certification cost. Within the three admission families, leave-one-family-out analysis admits with two and evaluates CB/OR certificates with the third. This diagnoses panel sensitivity and is distinct from the external validation of the full three-family gate. Folds are reported separately; overlapping admitted items are not treated as independent new observations. External-family transfer is evaluated once after all choices are frozen. Differences between selected populations are reported as selection effects, not as paired accuracy gains on an unchanged benchmark.

Visual relevance and image necessity. Language priors can confound apparent visual competence (Lee et al., 2025); we therefore distinguish whether an image is relevant to a question from whether it is needed to identify the target. The implemented visual check requires image–source alignment of 4/4 and question–image grounding of at least 2/4. The referent judgments use image descriptions and establish visual relevance. These checks do not by themselves establish image necessity. The image-removal and matched-replacement experiments assess sensitivity to visual input on the frozen evaluation set. A proposed stronger admission protocol would require grounding of at least 3/4, retain the alignment requirement of 4/4, and assess target identification directly from the original image. Two reviewers would independently record and lock their question-only judgments before viewing the image. Source titles, reference answers, and model outcomes would remain hidden. Each reviewer would assess whether the question alone admits at least two plausible targets and whether the image uniquely identifies the intended target. Agreement would determine the initial verdict; a senior reviewer would adjudicate disagreements using the same criteria. Images that do not resolve the target, targets already named in the question, and event descriptions that uniquely identify the target would fail this check. This stronger protocol is distinct from the admission procedure used for the reported collection.

The three admission families also attempt target recovery with the original image, no image, and a matched distractor, using identical question text and search budgets in fresh contexts. Any successful question-only recovery is a shortcut flag requiring rejection or revision and full recertification. Distractors must be plausible under the wording, depict a different target, and preserve image quality. The named-referent control isolates the contribution of target identification without supplying the requested fresh fact. All control outcomes, including failures with the original image, are retained. The external family tests the frozen set once; its results cannot be used to reselect that same evaluation cohort.

The release manifest freezes this gate and its panel before test-model evaluation. No evaluated model obtains a subset selected using its own image-ablation failures. Release-level reporting includes original-image successes, newly wrong and newly correct responses, and the fraction of original successes retained without the image. The resulting certificate is operational and budget-relative; neither a small average image effect nor model failure establishes universal necessity or impossibility.

A.2 Qualifications of the admission certificate

Equation 2 is an observed certificate relative to a declared panel and input conditions, not a claim that a fact is unknown in general or that a later search agent will obtain the supporting evidence. Four qualifications follow from that reading.

Abstention counts as an incorrect closed-book answer, so B_i certifies the absence of an *expressed* correct answer rather than the absence of the fact from a model's parameters. A model may still store the fact and refuse to emit it; the admission rule records only what was expressed under the closed-book prompt.

Repeated trials expose sampling variation within a model, not additional model families. The R trials per endpoint therefore bound decode-time instability for the declared panel; they do not enlarge $\mathcal{M}$.

The certificate is panel-relative. Whether Accept_i extends to models outside the admission panel is an empirical question, tested with families that had no role in screening (Section 5.1).

Finally, V_i establishes visual *relevance* and not image *necessity*. Relevance means that the visual referent is identifiable and the answer is not pixel-readable from the image. Necessity — that answering depends on seeing the image — is assessed separately: independent reviewers lock a text-only judgment before the image is revealed, and visual controls on the frozen release remove or replace the image (Sections 4.2 and 5.1).

These qualifications bound the question-level claim. Admission establishes an information gap for the panel and inputs used in Eq. 2; it does not establish what evidence a search agent will actually obtain in a with-search run.

B Short-Circuit Certification

Fix the candidate set, Visual relevance verdicts, grader, ordered trial outcomes, and release postprocessing. For $K = |\mathcal{M}|R$, admission is the conjunction of K CB-failure indicators and K OR-success indicators.

Proposition. Exhaustive and short-circuit certification produce identical released item IDs under these fixed inputs.

Proof. A short-circuit rejection observes a false conjunct in Equation 2, which also causes exhaustive rejection. A short-circuit acceptance evaluates every conjunct and finds each true. Conversely, every exhaustive rejection contains a false conjunct that short-circuit evaluation encounters before accepting. The admitted sets therefore agree, and identical deterministic postprocessing gives identical release sets. *QED*

The equivalence holds for fixed trial outcomes. Fresh API calls may produce different outcomes; a cached replay must preserve the outcomes and any shared random-number sequence.

Cost. For n Visual relevance survivors and n_{CB} CB check survivors, exhaustive execution uses $2Kn$ predictions. Block skipping evaluates all CB trials and skips OR after CB check rejection, costing $Kn + Kn_{\mathrm{CB}}$. Short-circuit execution counts only reached trials; every admitted item still needs $2K = 24$. For fixed trial costs c_j,

$$\mathrm{E}[\mathrm{Cost}] = \sum_j c_j \Pr(\text{trial } j \text{ is reached}). \tag{4}$$

Operational cost additionally includes token lengths, provider prices, generation, Visual relevance, held-out checks, human review, retries, and latency. Changing the generator or admission predicate changes the selected population and requires a separate yield–cost comparison.

C Model Roster and Execution Settings

C.1 Execution profile and controlled comparisons

Table 4 summarizes the paired inputs, shared ReAct search profile, and controlled retrieval settings. The search limits apply to an agent episode; retrieval depth specifies the number of results exposed in the controlled comparison.

Paired interventions. Retrieval-depth and reranking comparisons share a declared retrieval snapshot. Context comparisons preserve the question and gold evidence while varying the added passages; the irrelevant and conflicting additions are matched in length, source cues, and gold position. Paired accuracy changes are reported alongside coverage, since context selection can remove useful support as well as distractors. The intervention-selection example uses the same 3-result, 20-result, and 20-to-5 settings (Appendix G.1).

Table 4: **Evaluation inputs and reference execution settings.** The search profile applies to the common ReAct group in Table 5. Search-result counts and tool-action budgets are distinct: opening a page consumes an additional action.

Setting	Specification
Paired evaluation inputs	
CB	Image and question; no retrieval tools
WS	Image and question; access to the declared search tools
OR	Image, question, and gold evidence packet; no retrieval tools
Shared controls	Fixed item and model versions, answer-format instruction, and scoring rules; fresh context for each run
Shared ReAct search profile	
Search backend	Brave Web Search; English-language text queries
Available actions	Text search and page opening with a shared extractor
Results per search	Up to 20 ranked results
Action budget	At most 10 search/open actions per episode
Evidence budget	At most 16,000 visible text tokens, including snippets and extracted page text
Page limit	At most 2,000 extracted tokens per page
Time limit	120 seconds per episode
Generated search answers	Excluded
Controlled retrieval comparisons	
Retrieval depth	3, 10, and 20 results; the headline contrast is 20 versus 3
Reranking	Select 5 from the same 20-result pool, without access to gold answers

Models and conditions. Table 5 lists the twelve evaluation configurations. The paired intervention subset comprises claude-opus-5, qwen3.7-plus, Qwen3-VL-32B, and Qwen3-VL-8B. Construction-panel roles are specified separately in Appendix A. Each paired comparison fixes the question, item version, model configuration, and answer-format instruction. CB supplies the image and question without retrieval; OR additionally supplies the gold evidence packet; WS exposes the declared search tools. Conditions use fresh contexts, and final answers follow the same equivalence rules.

Table 5: **Evaluation model roster.** The shared group uses a common ReAct scaffold (Yao et al., 2023). Labels follow those used in the result tables.

Model display label	Evaluation group
`claude-opus-5`	Frontier reference
`gpt-5.5`	Frontier reference
`gemini-3.1-pro`	Frontier reference
`qwen3.7-plus`	Common ReAct
`GLM-5V-turbo`	Common ReAct
`qwen3.7-flash`	Common ReAct
`Qwen3-VL-32B`	Common ReAct
`qwen3-vl-235b`	Common ReAct
`Llama-4-Maverick`	Common ReAct
`Qwen3-VL-8B`	Common ReAct
`Qwen3-VL-30B-A3B`	Common ReAct
`Llama-4-Scout`	Common ReAct

Controlled comparisons. The retrieval-depth comparison exposes three versus twenty results; the reranking condition selects five from the same twenty-result pool without access to gold answers. Context interventions preserve the question and gold evidence while varying the added passages. Paired changes in accuracy and evidence coverage are reported together, since context selection can change both the availability and use of support.

Optional normalized score. When the measured CB–OR gap is positive,

$$\eta = \frac{\mathrm{Acc}_{\mathrm{WS}} - \mathrm{Acc}_{\mathrm{CB}}}{\mathrm{Acc}_{\mathrm{OR}} - \mathrm{Acc}_{\mathrm{CB}}}. \tag{5}$$

The ratio is undefined for a zero denominator and need not lie in $[0, 1]$. When selection makes CB close to zero and OR close to one, it largely duplicates WS accuracy; raw scores and paired changes remain primary.

Trace matching and unresolved support. Trace review inspects only content shown before the final answer. Matching preserves entity, relation, date, unit, and results scale; equivalent sources require a sufficiency judgment about the same fact. Bare URL hits, unopened pages, and spans lost to truncation do not establish availability. Independent review samples both matcher-positive and matcher-negative decisions, retaining false-positive, false-negative, and unresolved counts.

Let $A \in \{\mathrm{present}, \mathrm{absent}, \mathrm{unresolved}\}$, $p_j = \Pr(A = j)$, and $u_j = \Pr(C = 1 \mid A = j)$ for nonempty groups. Then

$$\mathrm{Acc}_{\mathrm{WS}} = p_{\mathrm{present}} u_{\mathrm{present}} + p_{\mathrm{absent}} u_{\mathrm{absent}} + p_{\mathrm{unresolved}} u_{\mathrm{unresolved}}. \tag{6}$$

With correct observed labels, coverage lies between p_{present} and $p_{\mathrm{present}} + p_{\mathrm{unresolved}}$ until unresolved cases are adjudicated. Empty-group conditional rates are undefined. The binary decomposition in the main text applies to resolved labels with an explicit denominator. Staleness, competing facts, and contradiction are additional, potentially overlapping tags; evidence–correctness cells locate an outcome without assigning its cause.

C.2 Changes across evidence conditions

The decomposition also clarifies the tradeoff algebraically. For two conditions on a shared cohort, let c_j, u_j, and v_j denote coverage, U_1, and U_0 in condition j. Subtracting their accuracies gives

$$\Delta\mathrm{Acc} = (c_2 - c_1)(u_1 - v_1) + c_2(u_2 - u_1) + (1 - c_2)(v_2 - v_1). \tag{7}$$

The first term is the contribution of changed coverage evaluated at the initial conditional rates; the other terms retain changes in conditional correctness. If both conditional rates stay fixed and $u_1 > v_1$, increasing coverage improves accuracy. A decrease in supported correctness can offset that benefit. Since the supported population may also change, these terms are descriptive contributions, not separate causal effects. Their value is to make the accompanying rate changes explicit when comparing retrieval conditions.

D Stratified Independent Review

The audit design samples 600 versioned questions from the frozen 6,000-item release, rather than treating sampling as a census. Five mutually exclusive domains receive proportional allocations based on manifest counts N_h. Selection is uniform without replacement within each domain. The records retain the sampling seed, inclusion probabilities, item versions, and source snapshots.

Three calibrated reviewers independently assess exact answer correctness, event unambiguity, image necessity, absence of pixel-only answers, supporting-span sufficiency, and temporal validity. Reviewers cannot assess their own items and are blinded to model identities and outcomes. Image relevance alone is insufficient: text-only judgments are locked before the image is revealed, testing whether it resolves an otherwise ambiguous referent. Reviewers extract the answer and supporting span before seeing the gold label. Majority labels, pre-discussion agreement, and independent adjudication remain separate.

A joint pass requires all six criteria to pass with complete records. Any confirmed failure makes an item a defect; remaining cases are unresolved or missing. These outcomes partition the original sample. Corrections create new versions; original outcomes remain visible and an independent follow-up sample evaluates the repaired release. Table 6 illustrates the structure; release validity is evaluated using the collected reviewer records.

Table 6: **Stratified human audit results.** Domain sizes, allocation, and outcomes are reported to demonstrate the quality of our benchmark. Joint pass requires all six criteria; pass, defect, and unresolved counts sum to n_h. Pass percentages retain unresolved items in the denominator.

Domain	N_h	n_h	Joint pass	Defect	Unresolved	Pass (%)
Science/technology	1,800	180	180	0	0	100.00
Sports	1,200	120	120	0	0	100.00
Politics	1,200	120	119	0	1	99.17
Culture	900	90	90	0	0	100.00
Business	900	90	90	0	0	100.00
Total	6,000	600	599	0	1	99.83

Let J_{hi} indicate joint pass, with zero assigned to unresolved or missing cases, and $W_h = N_h/N$. Population estimation uses

$$\widehat{p}_J = \sum_h W_h \overline{J}_h, \qquad \widehat{\mathrm{Var}}(\widehat{p}_J) = \sum_h W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{J,h}^2}{n_h}, \tag{8}$$

where $s_{J,h}^2$ is the within-domain sample variance. An unresolved-case sensitivity band extends from $\widehat{p}_J$ to $\widehat{p}_J + \sum_h W_h U_h / n_h$, separately from sampling uncertainty. Criterion-level and event-level estimates retain their own indicators and denominators; neither substitutes for joint item-level validity.

D.1 Audit sample-size analysis

We quantify how audit size affects the probability of detecting at least one defective item under specified defect rates and recognition probabilities. The calculations use uniform sampling without replacement as a reference; the domain-stratified audit follows the allocation described above. The calculations below describe uniform sampling as a reference design; the domain-stratified audit above uses its own inclusion probabilities. Including a defect in a sample and measuring the release's defect rate are different statistical objectives. For D defective items in a fixed population of N, a uniform sample of n includes at least one defect with probability

$$P_{\mathrm{detect}}(N, D, n) = 1 - \frac{\binom{N-D}{n}}{\binom{N}{n}}. \tag{9}$$

This formula assumes perfect recognition of a sampled defect. At $N = 6,000$ and a 1% defect rate, $n = 300$ gives 95.5% detection; at a 0.5% defect rate it gives 78.6%.

Table 7: **Finite-population audit design.** Probability (%) of including at least one defective item when sampling uniformly without replacement from a fixed population of 6,000 items. Columns assume exactly 30, 60, or 120 defective items and perfect recognition of a sampled defect.

Audited items	0.5% defects	1% defects	2% defects
100	39.7	63.7	87.0
200	63.9	87.1	98.4
300	78.6	95.5	99.8
500	92.7	99.5	> 99.9

With independent item-level recognition probability s, detection becomes

$$P_{\mathrm{detect}}(N, D, n; s) = 1 - \sum_{k=0}^{\min(D,n)} \frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}} (1-s)^k. \tag{10}$$

Here s characterizes the audit process rather than an individual reviewer. The calculation excludes false positives; shared misconceptions can violate independent recognition. At a 1% defect rate and $s = 0.8$, a 300-item audit detects at least one defect with probability 91.45%, and 364 items are needed for 95% detection. At $s = 0.6$, the requirement is 486.

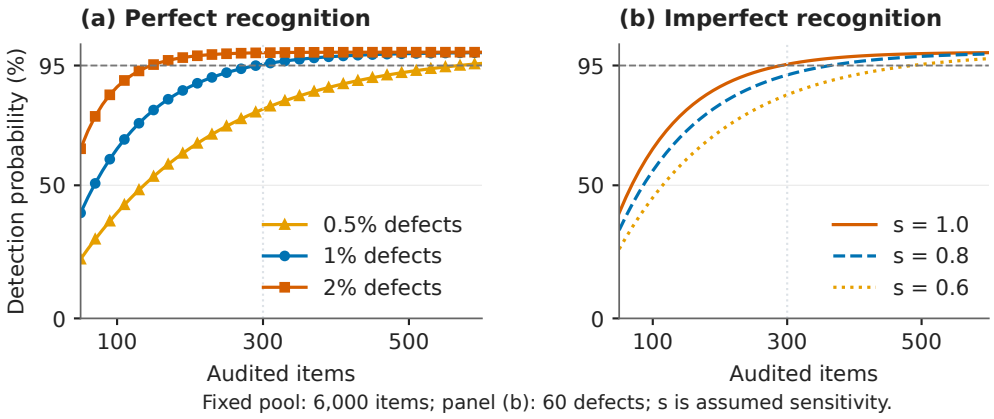

Figure 5: **Audit size and recognition sensitivity.** Exact detection under fixed defect counts (left) and alternative recognition probabilities at a 1% defect rate (right). Reference lines mark 95% detection and the 300-item sample.

Table 8: **Analytical sensitivity to imperfect recognition.** Fixed population 6,000; false-positive recognition is excluded and different sampled defects are recognized independently with probability s. The minimum integer sample size is obtained by exact hypergeometric summation. The recognition probability is a design parameter.

Defects (%)	Assumed s	Detection at $n = 300$ (%)	n for 95% detection
0.5	0.6	59.95	950
0.5	0.8	70.68	712
0.5	1.0	78.62	569
1.0	0.6	84.01	486
1.0	0.8	91.45	364
1.0	1.0	95.46	291
2.0	0.6	97.47	246
2.0	0.8	99.28	184
2.0	1.0	99.80	147

D.2 Paired estimation

Every interval uses 2,000 percentile resamples of the 30 whole builds, sharing build multiplicities across configurations and conditions. Events do not recur across builds in this design. Conditional rates are recomputed from resampled counts; averaging daily ratios would change the estimand. The complete baseline and joint cells follow.

Why resample already observed results? An accuracy difference is exact for the recorded cohort, but its sign or size may depend on the particular dates and events selected. Resampling approximates this sampling variation using the observed builds as the empirical distribution. For example, the gemini-3.1-pro-minus-qwen3.7-plus estimate is -0.45 pp with interval $[-1.18, 0.28]$: the evidence does not support a stable directional ordering at this interval level, although it also does not establish equivalence. The claude-opus-5-minus-gpt-5.5 contrast, 1.28 pp $[0.65, 1.88]$, has a positive interval under the same procedure. Keeping both models on the same resampled questions isolates uncertainty in their difference. The procedure estimates sampling uncertainty in paired score differences across the observed builds.

What a paired build resample does. One resample draws 30 indices with replacement from the 30 builds; for example, build 3 may occur twice while build 7 is absent. Both copies of build 3 contain

exactly the same item records and all compared conditions. The score or difference is recomputed from the resulting summed counts; conditional rates are not averages of build percentages. Repeating this procedure 2,000 times gives the empirical distribution used for the percentile interval. These are resamples of existing results, not new model runs or 2,000 newly collected builds. Pairing makes each contrast respond to the same question composition. The study assumes independent builds with dependence inside them; empirical events spanning dates must be clustered together, and the number of distinct clusters must be reported. These intervals quantify sampling variation across builds under the stated dependence assumptions; they do not account for provider drift or annotation uncertainty.

Table 9: **Complete three-condition results.** Twelve configurations on 30 builds of 200 items. All rates are percentages; intervals use 2,000 paired build resamples. Conditional rates use their joint-cell denominators.

Config.	CB	WS [95% interval]	OR	Cov	U_1	U_0
Frontier reference: native tool profile						
claude-opus-5	9.5	61.9 [59.9, 63.7]	94.4	74.0	80.9	7.6
gpt-5.5	10.0	60.6 [58.5, 62.6]	95.0	70.0	82.6	9.3
gemini-3.1-pro	8.4	57.6 [55.7, 59.4]	94.1	74.6	75.1	6.3
Common ReAct scaffold: shared tool interface and budget						
qwen3.7-plus	10.0	58.1 [55.9, 60.1]	93.9	66.9	81.9	9.7
GLM-5V-turbo	8.9	53.8 [51.9, 55.5]	93.5	71.9	71.9	7.5
qwen3.7-flash	11.2	53.8 [51.6, 56.0]	94.2	63.1	78.9	10.9
Qwen3-VL-32B	8.7	50.7 [48.9, 52.5]	93.5	73.1	66.5	7.9
qwen3-vl-235b	10.9	51.0 [48.9, 52.9]	92.7	66.1	71.9	10.2
Llama-4-Maverick	9.5	46.7 [44.8, 48.6]	92.7	67.1	65.5	8.3
Qwen3-VL-8B	10.4	47.1 [45.1, 49.0]	92.5	60.4	70.8	10.8
Qwen3-VL-30B-A3B	9.6	43.6 [41.6, 45.6]	91.8	67.8	59.9	9.4
Llama-4-Scout	11.7	44.5 [42.4, 46.5]	92.0	62.3	64.1	12.3

Table 10: **Complete evidence–correctness counts.** Each row totals 6,000. The last column is a percentage of all items and equals the contribution $(1 - \mathrm{Cov})U_0$ to WS; it is not the conditional rate U_0.

Config.	$C{=}1, G{=}1$	$C{=}1, G{=}0$	$C{=}0, G{=}1$	$C{=}0, G{=}0$	$P(C{=}1, G{=}0)$
claude-opus-5	3,594	118	847	1,441	1.97
gpt-5.5	3,468	167	729	1,636	2.78
gemini-3.1-pro	3,360	96	1,117	1,427	1.60
qwen3.7-plus	3,291	192	725	1,792	3.20
GLM-5V-turbo	3,100	126	1,213	1,561	2.10
qwen3.7-flash	2,989	241	800	1,970	4.02
Qwen3-VL-32B	2,916	128	1,468	1,488	2.13
qwen3-vl-235b	2,850	207	1,114	1,829	3.45
Llama-4-Maverick	2,639	163	1,388	1,810	2.72
Qwen3-VL-8B	2,568	257	1,057	2,118	4.28
Qwen3-VL-30B-A3B	2,437	181	1,630	1,752	3.02
Llama-4-Scout	2,396	277	1,343	1,984	4.62

The six adjacent contrasts in Table 11 retain the fixed configuration order and use intervals for the paired differences. claude-opus-5–gpt-5.5 and Qwen3-VL-30B-A3B–Llama-4-Scout exclude zero; the four middle intervals cross zero. These pointwise intervals do not provide a family-wise superiority test. The gemini-3.1-pro/qwen3.7-plus contrast spans the two tool profiles; the remaining five are within-group.

Each intervention partitions responses into both correct, both wrong, corrected, and newly wrong. Net change is $(n_{\text{corrected}} - n_{\text{newly wrong}})/N$. These discordant pairs also underlie the classical paired-proportion test (Dietterich, 1998); the reported intervals retain build clustering. For claude-opus-5 reranking, 80 corrected and 183 newly wrong responses give $100(80 - 183)/6000 = -1.7167$ percentage points.

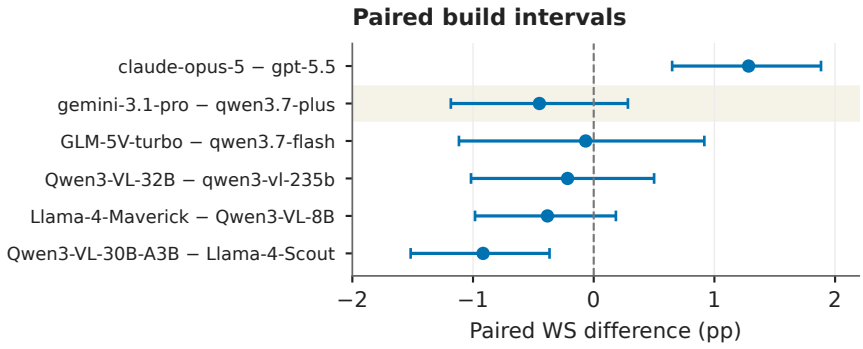

Figure 6: **Paired differences across configurations.** Point estimates and 95% build intervals preserve the shared-item comparison and fixed configuration order.

Table 11: **Descriptive paired comparisons.** A and B denote the first and second named configurations. Four transition cells sum to 6,000. Intervals resample complete builds and are pointwise, without multiplicity adjustment.

Contrast	Δ WS [95% interval]	A only	B only	Both correct	Both wrong
claude-opus-5 minus gpt-5.5	1.28 [0.65, 1.88]	175	98	3537	2190
gemini-3.1-pro minus qwen3.7-plus	-0.45 [-1.18, 0.28]	310	337	3146	2207
GLM-5V-turbo minus qwen3.7-flash	-0.07 [-1.12, 0.92]	325	329	2901	2445
Qwen3-VL-32B minus qwen3-vl-235b	-0.22 [-1.02, 0.50]	236	249	2808	2707
Llama-4-Maverick minus Qwen3-VL-8B	-0.38 [-0.98, 0.18]	226	249	2576	2949
Qwen3-VL-30B-A3B minus Llama-4-Scout	-0.92 [-1.52, -0.37]	168	223	2450	3159

E Visual Controls

The visual-control study uses the same 6,000 frozen questions for four selected configurations: claude-opus-5, gpt-5.5, gemini-3.1-pro, and Llama-4-Maverick. The primary answering comparison uses the original image, no image, and text naming the correct referent without revealing the requested fact. A separate target-identification probe disables search, and the OR reference supplies gold evidence with the original image. Image replacement is treated separately because changing the pictured referent can change the correct answer.

Each primary input condition contains 24,000 item–configuration observations over 6,000 unique questions. All conditions retain the same item identifiers, including questions answered incorrectly with the original image. The question, event date, scoring rules, and within-mode tool budgets remain fixed. Inputs without an image exclude image-derived captions, OCR, filenames, and other metadata. The named-referent input supplies only the identity, without the requested fact, supporting passage, or suggested search query.

E.1 Admission records and evaluation outcomes

The original visual-grounding threshold is 2 on the 0–4 rubric, corresponding to an identifiable referent. It does not itself establish that the question cannot be answered without the image. The visual controls test this additional property on the frozen release; they are not used to select a different subset for each evaluated configuration.

Table 12: **Complete paired intervention results.** Changes are in percentage points, with 95% paired build intervals. The first two columns change oracle context; the last two change search retrieval. All contrasts share 6,000 items.

Config.	Irrel. vs clean	Conflict vs irrel.	$k = 20$ vs 3	Rerank vs 20
claude-opus-5	-0.5 [-0.7, -0.3]	-10.1 [-10.8, -9.3]	+11.0 [+10.1, +11.9]	-1.7 [-2.3, -1.2]
qwen3.7-plus	-0.9 [-1.1, -0.8]	-6.9 [-7.6, -6.3]	+11.8 [+10.9, +12.8]	-2.0 [-2.5, -1.6]
Qwen3-VL-32B	-1.3 [-1.6, -1.0]	-14.0 [-14.7, -13.2]	-1.4 [-2.4, -0.2]	+7.5 [+6.7, +8.3]
Qwen3-VL-8B	-0.2 [-0.5, -0.0]	-13.4 [-14.2, -12.6]	+6.7 [+5.6, +7.7]	+4.1 [+3.5, +4.8]

The CB and gold-evidence admission predicates concern the construction panel and its recorded screening trials. An admitted question has no correct answer in its prescribed CB screening calls and correct answers in all prescribed gold-evidence calls. These predicates describe observed screening outcomes, rather than universal guarantees about other configurations or later runs.

Evaluation responses are kept separate from the screening records. Fresh runs of a construction-panel configuration and runs of other configurations must also be identified separately. Reusing the screening responses would reproduce the admission outcomes by definition; such a replay is not an independent evaluation. The named-referent control changes the available identity information and is not constrained by the CB predicate on the original input.

Residual CB successes should be inspected for answer leakage, ambiguity, and grading inconsistencies, while allowing for correct answers obtainable from a model's prior knowledge or inference. Their presence alone does not identify which explanation applies. Likewise, OR failures require checking both evidence sufficiency and the evaluated model's response. Any correction that changes the question set must receive a new release version.

E.2 Complete results and conditional retention

For configuration m and WS control j, let $n_{ab}^{(m,j)}$ count questions with original-image correctness a and control correctness b. With $N = 6{,}000$ questions per configuration, control accuracy and conditional retention are

$$A_{m,j} = \frac{n_{11}^{(m,j)} + n_{01}^{(m,j)}}{N}, \qquad R_{m,j} = \frac{n_{11}^{(m,j)}}{n_{11}^{(m,j)} + n_{10}^{(m,j)}}.$$

Accuracy includes newly correct answers, whereas retention conditions on original-image success. Consequently, the accuracy ratio $A_{m,j}/A_{m,\text{full}}$ can exceed one, while retention cannot. Aggregate accuracies average equally sized configuration samples; aggregate retention pools the underlying retained and original-success counts. Table 13 summarizes the conditions, and Table 14 reports each configuration separately. The original-image rows are the main-evaluation runs of Table 2: this condition has exactly the inputs of the main evaluation, so it is reused rather than repeated. Only the no-image and named-referent inputs, and the target-identification probe, are specific to this study.

Table 13: **Visual-control results.** Four configurations each receive the same 6,000 questions. CB and OR describe evaluation runs, not admission trials. Accuracies average equally sized configuration samples; WS retention and the accuracy ratio pool their underlying counts. Target identification is measured separately without search. All numeric entries are percentages. OR uses the original image and gold evidence; other OR inputs are not evaluated.

Input	Target identification	CB	WS	WS retained	WS accuracy ratio	OR
Original image	97.0	9.4	56.7	100.0	100.00	94.1
No image	18.0	3.5	17.0	25.0	29.99	—
Named referent	Supplied	11.8	73.0	95.0	128.78	—

Table 14: **Visual-control results by configuration.** Each configuration and input uses the same 6,000 questions, giving 24,000 item–configuration observations per input. The original-image rows reuse the main-evaluation runs of Table 2, since that condition has identical inputs; the other rows are specific to this study. Target identification is a separate probe without search; *Supplied* denotes an input identity, not a recognition score. CB and WS report answer accuracy, and retention refers to original-image WS successes; pooling across configurations gives 25.0% without the image and 95.0% with the named referent. OR supplies oracle evidence with the original image. All numeric entries are percentages.

Configuration	Input	Target identification	CB	WS	WS retained	OR
claude-opus-5	Original	98.0	9.5	61.9	100.00	94.4
	No image	15.0	3.2	18.0	22.49	—
	Named referent	Supplied	12.7	77.0	95.93	—
gpt-5.5	Original	97.0	10.0	60.6	100.00	95.0
	No image	17.0	3.8	17.5	25.01	—
	Named referent	Supplied	12.5	75.0	94.99	—
gemini-3.1-pro	Original	97.0	8.4	57.6	100.00	94.1
	No image	19.0	3.2	16.5	26.04	—
	Named referent	Supplied	10.5	71.0	93.98	—
Llama-4-Maverick	Original	96.0	9.5	46.7	100.00	92.7
	No image	21.0	3.8	16.0	26.80	—
	Named referent	Supplied	11.4	69.0	95.00	—

E.3 Paired outcomes and interpretation

Table 15 retains both losses and gains under each WS control. Across configurations, the original input produces 13,605 correct answers. Removing the image retains 3,395 of these successes and adds 685 newly correct answers. Retention is therefore $3{,}395/13{,}605 = 25.0\%$, whereas accuracy is $4{,}080/24{,}000 = 17\%$. The accuracy ratio of 29.99% includes the new successes and must not be substituted for retention.

Table 15: **Paired WS outcomes.** Each configuration row contains 6,000 questions; pooled rows contain 24,000 item–configuration observations. The first two numeric columns sum to the original correct count, and the first and third sum to the control correct count.

Configuration	Control	Both correct	Correct to wrong	Wrong to correct	Both wrong
claude-opus-5	No image	835	2,877	245	2,043
	Named referent	3,561	151	1,059	1,229
gpt-5.5	No image	909	2,726	141	2,224
	Named referent	3,453	182	1,047	1,318
gemini-3.1-pro	No image	900	2,556	90	2,454
	Named referent	3,248	208	1,012	1,532
Llama-4-Maverick	No image	751	2,051	209	2,989
	Named referent	2,662	140	1,478	1,715
Pooled	No image	3,395	10,210	685	9,710
	Named referent	12,924	681	4,596	5,799

The named-referent condition retains 12,924 original successes and gains 4,596, yielding 95% retention and 73% accuracy. These gains explain why its accuracy ratio exceeds 100%. The loss of 75% of original successes after image removal describes a conditional set of model responses; it is not the proportion of all questions for which images have been proved universally necessary.

E.4 VISUAL CONTROLS VALIDATE THE ROLE OF TARGET IDENTIFICATION

The visual controls assess a property that the admission checks alone do not establish: whether the evaluated systems depend on the image to identify what they should search for. The original grounding threshold requires an identifiable visual referent, but does not rule out successful answering from textual cues. Removing the image preserves only 24.9% of original WS successes, whereas supplying the correct referent preserves 95%. Together with the independent target-identification probe, this contrast supports the intended role of visual information in resolving the search target.

The named-referent condition further separates access to the target identity from access to the requested fact. Supplying the identity raises with-search accuracy by 16.3 points, while closed-book accuracy rises by only 2.4 points. Resolving the referent therefore recovers search successes without making the dated fact readily answerable from the model's existing knowledge. This control provides a useful reference for investigating failures associated with target identification and for evaluating search when the correct identity is available.

These comparisons connect question admission to observed system behavior. They validate the intended task structure on the same frozen questions across the selected configurations, rather than assuming that passing a visual-grounding check guarantees image dependence. The benchmark therefore supports evaluation of both the contribution of visual target identification and the continuing need for external factual evidence.

F VERSIONED RELEASE LIFECYCLE

Versioned cohorts. A cohort fixes build and item versions, evaluation window, and tool profile. Comparisons within it preserve item pairing. Between-build variation measures sensitivity to the selected questions rather than calibrating different builds to a shared difficulty scale. The lifecycle progresses through proposal, automatic admission, independent review, evaluation, archival, and correction or retirement. Review thresholds, unresolved-label rules, and repair procedures are fixed in the release manifest.

Source loss, altered answerability, or a disputed label triggers review. Confirmed defects create explicit supersession records while historical scores remain attached to their original versions. Public metadata and authorized assets are distinguished from restricted archives by provenance and access records. Live search and replay retain separate identities.

Build manifests and fixed item versions. To document renewal, we specify a build manifest that links each dated cohort to its source material and construction configuration. Each entry records the build identifier and timestamp, the publication range of the released items' sources, the numbers, admitted information, and released questions, and the configuration version. The corresponding item-version index maps each released question to its build, publication timestamp, and retained support record. Corrections create new item versions; previous evaluations remain attached to the versions on which they were run. Table 16 illustrates this reporting format with corresponding entries. CFG-E1 denotes the construction configuration used throughout the 30 dated builds, including the visual-relevance check, three-model panel, and four closed-book and four oracle-evidence trials per question. For each item, freshness is determined by comparing its source publication timestamp with the timestamp of the build in which the item was made.

Renewal persistence across builds. Section 5.1 reports human audit, external closed-book and oracle-evidence rates, and visual controls on the pooled 6,000-item release. Those summaries establish that the released questions satisfy the task definition; they do not by themselves show that the same definition held in each dated build. The build manifest already records the quantities needed for that check. Admission yield is the admitted count divided by the candidate count. Across the thirty dated builds the yield ranges from 40.7% (B28: $242/594$) to 55.8% (B30: $276/495$), and the thirty-build total is 47.9% ($7{,}861/16{,}401$). Every build admits more than the 200-item release cap, so no dated cohort was completed by relaxing Eq. 2. Table 18 reports the yield of each build; Table 16 records the corresponding source windows.

The same item-version index supports a second check that requires no new model call: group the already-collected evaluation runs by build and report, for the six systems unrelated to the admission

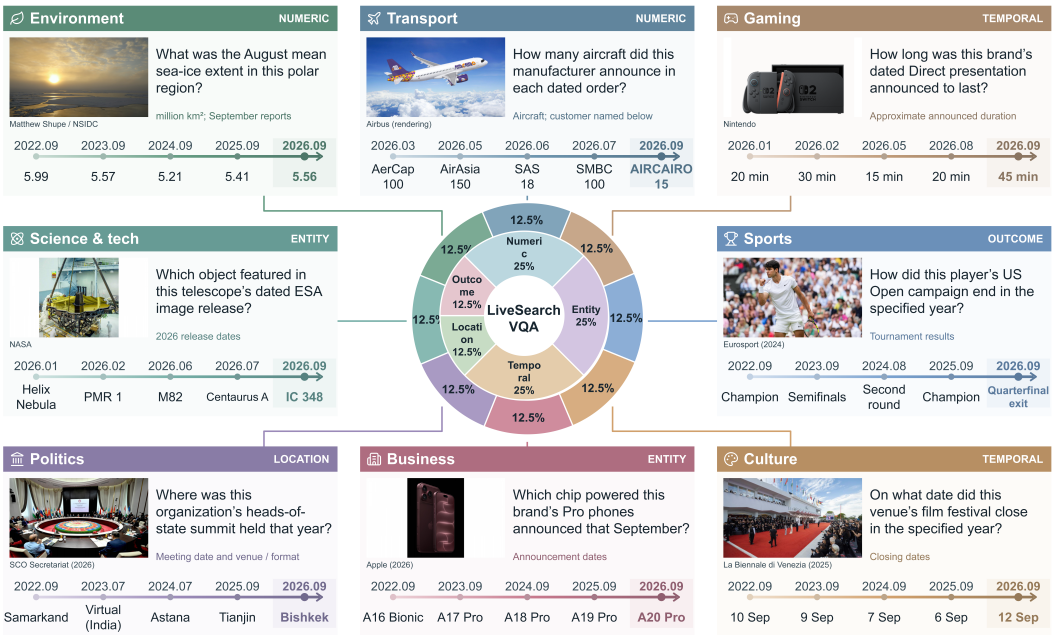

Figure 7: **Benchmark overview and dated question families.** LIVESEARCHVQA comprises 6,000 multimodal search questions grounded in visual subjects and dated external evidence. The outer ring groups questions by domain and the inner ring by answer form; these are cross-cutting labels, and the percentages describe the 40 illustrated instances rather than the full benchmark. Each timeline shows five dated questions, with darker arrows toward September 2026; earlier versions retain their answers.

panel, the mean closed-book and oracle-evidence accuracy of that build. Across the thirty builds those systems remain in 8.6–10.8% closed book and 92.0–95.0% oracle evidence, with thirty-build means of 9.7% and 93.6% (Table 19), matching the pooled rates in Section 5.1. A large drift in either rate would mean that later builds were easier to answer from prior knowledge, or harder to recover from the recorded packet, than earlier ones. The stable yield and the absence of such drift are the operational meaning of renewal persistence: each new batch is admitted under the same rule, not averaged into a mixed-quality pool.

G VALIDATION PROTOCOLS

The following studies specify the validation of the benchmark and diagnostic comparisons.

Visual contribution. The comparison evaluates each frozen question with its original image, no image, a matched distractor, and text naming the intended referent without revealing the answer. Question wording, event date, tools, and budget remain fixed. Distractors are reviewed for answer leakage and ambiguity. Paired gains, losses, coverage, and correctness are reported by visual anchor and grounding score. A separate frozen panel selects any image-dependent subset before the evaluation roster is run.

Held-out certificate transfer. Panel-external models receive fresh $R = 4$ CB and OR trials under the same profile. Define $B = 1$ for no CB success and $E = 1$ for all OR successes. The complete 2×2 table reports the joint transfer probability. Marginals alone imply only

$$\max(0, \Pr(B) + \Pr(E) - 1) \le \Pr(B \wedge E) \le \min(\Pr(B), \Pr(E)). \tag{11}$$

Fresh construction-panel and held-out results are reported separately, with family overlap and repetition counts explicit. For the new protocol, $K_E = R|\mathcal{E}|$ gives four trials per condition for one external family and eight for two; this study uses twelve trials from three panel models. Joint-pass rates require the full conjunction under the declared K_E, and remain separate from single-trial rates.

Filter contribution and selection. Joint admission is a design choice: it lowers yield relative to any single predicate. The value of that choice is the quality of the selected set, not a paired accuracy gain on a fixed item list. The comparison therefore holds one cached candidate pool fixed and changes only the admission gate. Four gates are defined on the same visually screened pool:

$$V, \qquad V \wedge B, \qquad V \wedge O, \qquad V \wedge B \wedge O,$$

where V, B, and O are the predicates of Eq. 2. An optional fifth gate uses a single admission endpoint in place of the full panel, to separate the contribution of the conjunction from the contribution of replicating it. Generation, scoring, and the held-out evaluation roster stay fixed.

Each gate produces its own admitted set. The quality of a set is the rate at which its items reproduce both closed-book failure and supplied-evidence success for model families outside the admission panel. Because the four sets are different populations, differences in that rate are selection effects: they measure which questions a gate keeps, not how the same questions would have scored under a different rule. Reported quantities are therefore the admitted count, the held-out joint-pass rate with a resampling interval, the four $B \times O$ cells on the admitted set, and the cost in admission trials per released item. Rankings, evidence–correctness profiles, answer types, source domains, and evidence length are secondary descriptors of the same sets. Visual relevance, evidence-first generation, and feedback memory are not part of this comparison; they change which candidates are written, and are ablated separately.

The released 6,000-item stream cannot be reused for this comparison. Admission on that stream is short-circuited: a candidate that fails B does not receive oracle-evidence trials (Appendix B). The cell required to form $V \wedge O$ is therefore missing for every closed-book rejection. Completing the four gates on the release stream would require fresh oracle-evidence trials on those rejected candidates. The comparison belongs on a candidate stream that records every trial for every visually screened item, disjoint from the frozen release. A probability sample of items rejected by $V \wedge B \wedge O$ but admitted by a weaker gate estimates false rejection under the joint rule.

Context and retrieval interventions. Clean gold evidence is compared with gold plus five irrelevant passages and gold plus five conflicting passages. The added contexts match token length, source cues, and gold position; human checks verify support preservation and the intended conflict. Both corrected and newly wrong answers are retained. Retrieval depths 3, 10, and 20 share a declared retrieval interface and snapshot. A gold-blind reranker selects five passages from the top twenty, with its identity, cost, support retention, context length, and WS accuracy recorded.

Comparison and renewal. Compatible subsets of prior benchmarks retain their original task definitions and use matched model versions, tools, budgets, and graders. Comparisons examine no-search answerability, gold-context answerability, and independent validity alongside task scope and source dates. Prospective renewal analysis records timestamps, release counts, shortfalls, unique events, duplicate removals, domains, answer types, and per-build outcomes. Events spanning builds are grouped across dates; generator or rule updates are evaluated in a new event-disjoint period.

G.1 Do diagnostics guide effective interventions?

Setup and selection rules. We illustrate the selection protocol with an experiment using the four configurations and intervention definitions in Table 12. The split assigns 24 builds (4,800 items) to development and six builds (1,200 items) to testing, with events confined to one split. Let A retrieve three results, B retrieve twenty, and C rerank the same twenty-result pool to five without access to gold answers. Both selectors use development runs under A and B; the effect of C and all test outcomes remain hidden until their choices are fixed. The score-only rule selects B when its development accuracy exceeds A and otherwise selects C. The diagnostic rule selects C when coverage increases but supported correctness U_1 falls by at least two percentage points; otherwise it selects B. This threshold and rule are fixed for all four configurations. The evaluator then compares paired test outcomes under A, B, and C.

Selection outcomes. Table 20 connects development diagnostics to intervention choices and their test outcomes. For Qwen3-VL-8B, development accuracy increases by 6.79 points under B, while coverage rises by 18.75 points and U_1 falls by 6.03. The score-only rule retains B, whereas the

diagnostic rule recommends C, increasing the test gain from 6.33 to 10.17 points in this example. The selectors agree for the other three configurations, so the additional benefit comes from this 8B decision. Across configurations, diagnostic selection yields a mean gain of 9.35 points, compared with 8.40 for the score-only rule and 8.44 for shared C. The respective differences of 0.96 and 0.92 points are computed before rounding.

Implication for benchmark value. Table 20 illustrates how LIVESEARCHVQA's evidence records can turn an improving search score into a testable recommendation for further improvement. The same diagnostic rule recommends retaining deeper retrieval for some systems and adding reranking for others. Because the supported subsets can change, a fall in U_1 motivates an intervention test without identifying its cause. The selection rule and its two-point threshold are fixed in advance on development builds, so this example demonstrates that the diagnosis yields an actionable rule, not that the rule is optimal. With four configurations and a single decisive disagreement, the mean difference of 0.96 points characterizes this comparison rather than an expected gain.

G.2 VALIDATING SUPPORT-STATE DIAGNOSIS WITH RETAINED RECORDS

Target and evaluation unit. The evaluation treats one frozen item–configuration WS run as the diagnostic unit and conditions on an incorrect final answer, $C = 0$. The target is whether the content actually visible before that answer contains sufficient support for the reference fact. A supported error has $G = 1$; an unsupported error has $G = 0$. The support label concerns the correct reference fact, rather than justification for the agent's incorrect answer. Equivalent sources count when they establish the same entity, relation, event or date, and value. Bare URLs, unopened pages, and passages removed by truncation cannot establish visibility.

Independent reference labels. This study samples incorrect WS runs across the frozen builds and evaluation configurations, retaining item, event, build, and model identifiers. Three reviewers independently inspect the archived agent-visible content against the reference fact and source evidence, without seeing either diagnostic condition's predictions. Disagreements receive independent adjudication. Cases that remain unresolved are reported separately and excluded before fixing the common resolved-case denominators for both conditions; they are not relabeled as unsupported. Sampling counts and inclusion probabilities are recorded so that balanced diagnostic accuracy can be distinguished from the prevalence of the two failure states in the benchmark. This run-level evaluation is separate from the question-validity audit in Appendix D.

Paired information ablation. The trace-only condition receives the image, dated question, reference answer and accepted aliases, final response, and the exact pre-answer trace, including visible snippets, opened-page text, tool actions, and truncation boundaries. The record-aware condition receives the same inputs plus the retained support passage, its archived source context, publication timestamp, source identifier, and span offsets. The additional record specifies what evidence would be sufficient; it does not count as evidence that reached the original agent. A positive diagnostic decision must cite support inside the original visible trace. Both conditions exclude existing matcher verdicts, reviewer labels, and construction-panel outcomes. One frozen diagnostic judge, decision rubric, decoding configuration, and output schema are used in fresh contexts for both conditions, with enough input capacity to preserve the same complete trace. The schema requires a binary support prediction and an accompanying trace reference or explanation. Prompts and any calibration choices are set fixed on event-disjoint development data. Neither condition conducts new search or regenerate the agent's answer.

Metrics. Let G^{ref} denote the independently adjudicated support label and $\widehat{G}$ the diagnostic prediction, always within the cohort $C = 0$. Treat $G^{\mathrm{ref}} = 1$ as positive. Thus, TP and FN partition supported errors, while TN and FP partition unsupported errors. Define

$$\mathrm{BA} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right), \qquad \mathrm{BER} = 1 - \mathrm{BA}. \tag{12}$$

The relative reduction in balanced diagnostic error is

$$\mathrm{RER} = \frac{\mathrm{BER}_{\mathrm{trace}} - \mathrm{BER}_{\mathrm{record}}}{\mathrm{BER}_{\mathrm{trace}}}. \tag{13}$$

An experimental test. In this experiment, there are 600 resolved error cases with 300 in each reference support state. Table 21 specifies two confusion matrices on that common cohort. The trace-only recalls are $234/300 = 78\%$ for supported errors and $246/300 = 82\%$ for unsupported errors. The record-aware recalls are $276/300 = 92\%$ and $282/300 = 94\%$. Their averages are 80% and 93%, giving

$$\frac{(1 - 0.80) - (1 - 0.93)}{1 - 0.80} = \frac{0.20 - 0.07}{0.20} = 0.65. \tag{14}$$

The absolute balanced-accuracy difference is 13 percentage points. The 65% value is the relative reduction in balanced error. Because the two reference classes are equally sized in this example, ordinary diagnostic error also falls from $120/600$ to $42/600$; that equality need not hold under a different class distribution.

Interpretation and empirical reporting. The results show how a change in diagnostic reliability can coexist with unchanged search answers. A comparison retains paired case-level predictions, report both corrected and newly introduced diagnostic errors, and give class-specific recalls alongside BA and BER. Uncertainty is estimated by 2,000 paired cluster-bootstrap resamples, preserving both diagnostic conditions and all configurations for each sampled build. Events spanning builds are grouped into common clusters, and the number of represented clusters is reported. The same frozen protocol applies to later dated cohorts, which is how we intend to track whether diagnostic reliability persists under renewal.

H Additional related work

Table 1 compares documented task scope and controls; it does not report a controlled performance comparison across benchmarks. The following connections complete that positioning.

Knowledge-intensive visual reasoning. A-OKVQA, Encyclopedic VQA, and ViQuAE examine world knowledge and fine-grained entities (Schwenk et al., 2022; Mensink et al., 2023; Lerner et al., 2022); MultiModalQA combines text, tables, and images (Talmor et al., 2021). These tasks motivate separating referent identification from the acquisition of a fresh fact.

Renewal and benchmark exposure. LiveBench and LiveCodeBench collect new material; Dynabench uses human–model interaction (White et al., 2025; Jain et al., 2025; Kiela et al., 2021). Contamination studies caution that exposure is benchmark- and model-specific (Golchin & Surdeanu, 2024; Sainz et al., 2023). Our dated builds and retained CB trials support this assessment; publication time alone cannot establish model ignorance.

Tools and information discovery. WebGPT, ReAct, Self-Ask, and Toolformer develop browsing and tool use (Nakano et al., 2021; Yao et al., 2023; Press et al., 2023; Schick et al., 2023). Mind2Web, WebArena, VisualWebArena, and WebVoyager evaluate web interaction (Deng et al., 2023; Zhou et al., 2024; Koh et al., 2024; He et al., 2024). GAIA and AssistantBench cover practical tasks (Mialon et al., 2023; Yoran et al., 2024a); BrowseComp, WebWalker, and MM-BrowseComp stress information discovery (Wei et al., 2025; Wu et al., 2025; Li et al., 2025). Our admission and trace records can accompany these agents to diagnose the evidence behind an outcome.

Retrieval and assessment. RAG, Self-RAG, and BERGEN study answering with retrieved evidence (Lewis et al., 2020; Asai et al., 2024; Rau et al., 2024). Distraction and knowledge-conflict studies motivate controlled evidence interventions (Shi et al., 2023; Wang et al., 2024). G-Eval and Constitutional AI use structured model judgments and explicit criteria (Liu et al., 2023; Bai et al., 2022); our independent review assesses whether automated admission satisfies the benchmark's task definition.

Table 16: **Build manifest.** Each dated build links its released items to the source material and construction configuration that produced them. All timestamps are UTC in 2026; rows show an excerpt and the totals row covers all 30 builds. Admitted counts exceed the 200-item release cap in every build, so no release was completed by relaxing an admission threshold.

Build	Build time (UTC)	Source publication range (UTC)	Cand.	Admit.	Release	Config.
B01	08-21 12:00	08-20 00:00 to 08-21 11:00	460	231	200	CFG-E1
B02	08-22 12:00	08-20 17:00 to 08-22 10:00	497	248	200	CFG-E1
B03	08-23 12:00	08-21 22:00 to 08-23 09:00	534	265	200	CFG-E1
B04	08-24 12:00	08-22 17:00 to 08-24 09:00	520	262	200	CFG-E1
B05	08-25 12:00	08-23 22:00 to 08-25 10:00	550	269	200	CFG-E1
B06	08-26 12:00	08-24 14:00 to 08-26 08:00	576	274	200	CFG-E1
B07	08-27 12:00	08-26 05:00 to 08-27 11:00	493	254	200	CFG-E1
B08	08-28 12:00	08-26 16:00 to 08-28 09:00	612	276	200	CFG-E1
B09	08-29 12:00	08-28 00:00 to 08-29 08:00	528	263	200	CFG-E1
B10	08-30 12:00	08-28 13:00 to 08-30 10:00	562	258	200	CFG-E1
B11	08-31 12:00	08-29 20:00 to 08-31 07:00	618	275	200	CFG-E1
B12	09-01 12:00	08-31 07:00 to 09-01 09:00	480	251	200	CFG-E1
B13	09-02 12:00	08-31 18:00 to 09-02 11:00	539	257	200	CFG-E1
B14	09-03 12:00	09-01 15:00 to 09-03 10:00	587	271	200	CFG-E1
B15	09-04 12:00	09-03 03:00 to 09-04 08:00	513	261	200	CFG-E1
B16	09-05 12:00	09-03 19:00 to 09-05 09:00	569	268	200	CFG-E1
B17	09-06 12:00	09-04 13:00 to 09-06 10:00	633	276	200	CFG-E1
B18	09-07 12:00	09-06 08:00 to 09-07 07:00	490	255	200	CFG-E1
B19	09-08 12:00	09-06 21:00 to 09-08 11:00	554	271	200	CFG-E1
B20	09-09 12:00	09-08 01:00 to 09-09 09:00	525	266	200	CFG-E1
B21	09-10 12:00	09-08 14:00 to 09-10 10:00	591	264	200	CFG-E1
B22	09-11 12:00	09-10 06:00 to 09-11 08:00	469	249	200	CFG-E1
B23	09-12 12:00	09-10 16:00 to 09-12 11:00	606	265	200	CFG-E1
B24	09-13 12:00	09-11 23:00 to 09-13 07:00	542	256	200	CFG-E1
B25	09-14 12:00	09-12 17:00 to 09-14 10:00	578	267	200	CFG-E1
B26	09-15 12:00	09-14 04:00 to 09-15 09:00	508	260	200	CFG-E1
B27	09-16 12:00	09-14 20:00 to 09-16 08:00	547	272	200	CFG-E1
B28	09-17 12:00	09-15 15:00 to 09-17 09:00	594	242	200	CFG-E1
B29	09-18 12:00	09-16 20:00 to 09-18 08:00	631	259	200	CFG-E1
B30	09-19 12:00	09-17 13:00 to 09-19 07:00	495	276	200	CFG-E1
All 30-dated builds			16,401	7,861	6,000	CFG-E1

Rows show an excerpt; totals cover all 30 builds. *Definitions.* Candidate counts precede joint admission; admitted counts pass the visual/CB/OR checks; release counts enter the frozen cohort after selection. Each row links to the item-version index through its build ID.

Table 17: **Dataset and comparison units.** Publication freshness is checked when building an item; its event-specific meaning and historical evaluation remain fixed thereafter.

Benchmark property	Definition
Collection	6,000 questions; 30 dated builds of 200
Freshness at construction	Source published within the preceding 48 hours
Question record	Image, dated question, answer, source span, version
Admission profile	Visual relevance; three panel models; four CB and four OR trials each
Evaluation unit	Frozen item versions, shared window and tool profile
Answer forms	Numeric, temporal, entity, location, outcome

Table 18: **Admission yield by dated build.** Yield is admitted candidates divided by candidates before joint admission. Every build exceeds the 200-item release cap.

Build	Cand.	Admit.	Yield	Released
B01	460	231	50.2%	200
B02	497	248	49.9%	200
B03	534	265	49.6%	200
B04	520	262	50.4%	200
B05	550	269	48.9%	200
B06	576	274	47.6%	200
B07	493	254	51.5%	200
B08	612	276	45.1%	200
B09	528	263	49.8%	200
B10	562	258	45.9%	200
B11	618	275	44.5%	200
B12	480	251	52.3%	200
B13	539	257	47.7%	200
B14	587	271	46.2%	200
B15	513	261	50.9%	200
B16	569	268	47.1%	200
B17	633	276	43.6%	200
B18	490	255	52.0%	200
B19	554	271	48.9%	200
B20	525	266	50.7%	200
B21	591	264	44.7%	200
B22	469	249	53.1%	200
B23	606	265	43.7%	200
B24	542	256	47.2%	200
B25	578	267	46.2%	200
B26	508	260	51.2%	200
B27	547	272	49.7%	200
B28	594	242	40.7%	200
B29	631	259	41.0%	200
B30	495	276	55.8%	200
All 30 builds	16,401	7,861	47.9%	6,000

Table 19: **External closed-book and oracle-evidence rates by build.** Each cell averages the six systems unrelated to the admission panel on the 200 released items of that build.

Build	External CB (%)	External OR (%)
B01	9.5	93.9
B02	10.2	93.4
B03	9.7	94.2
B04	9.8	93.6
B05	10.3	92.9
B06	8.9	94.5
B07	9.5	94.0
B08	10.6	92.5
B09	9.2	94.1
B10	10.0	93.3
B11	9.7	92.7
B12	8.7	94.8
B13	9.9	93.7
B14	10.4	93.0
B15	9.3	94.3
B16	9.8	93.5
B17	10.8	92.0
B18	8.8	94.9
B19	9.6	93.8
B20	9.1	94.4
B21	10.2	92.8
B22	8.6	95.0
B23	10.5	92.2
B24	9.4	93.8
B25	10.0	93.1
B26	9.3	94.0
B27	9.7	93.5
B28	10.2	93.0
B29	10.4	92.5
B30	8.9	94.6
All 30 builds	9.7	93.6

Table 20: **Diagnostic-guided intervention selection.** All changes are in percentage points. Diagnostic choices are B, B, C, C; score-only choices are B, B, C, B. Test gains are relative to the same A condition, and means weight configurations equally.

	Development: $B - A$		Chosen with diagnosis	Test gain over A			
Configuration	ΔCov	ΔU_1		B	C	Diagnostic	Score-only
claude-opus-5	+15.00	-0.28	B	+10.67	+9.17	+10.67	+10.67
qwen3.7-plus	+16.67	-0.56	B	+11.33	+9.17	+11.33	+11.33
Qwen3-VL-32B	+15.00	-12.93	C	-1.83	+5.25	+5.25	+5.25
Qwen3-VL-8B	+18.75	-6.03	C	+6.33	+10.17	+10.17	+6.33
Mean	–	–	–	+6.63	+8.44	+9.35	+8.40

Table 21: Diagnosis on a fixed cohort of 600 incorrect WS cases, comprising 300 supported and 300 unsupported errors. Supported errors are the positive class. Both rows refer to the same cases, original search traces, and final answers; only the diagnostic inputs differ. BA is balanced accuracy and BER is balanced error rate.

Diagnostic input	TP	FN	FP	TN	BA (%)	BER (%)
Traces only	234	66	54	246	80.0	20.0
Traces + support records	276	24	18	282	93.0	7.0

Table 22: **Documented designs of closely related benchmarks.** LIVESEARCHVQA links versioned source evidence and joint admission records to paired evaluation and evidence traces. The comparison concerns benchmark design and scope; it does not report comparative model performance.

Benchmark	Data and renewal	Construction controls	Evaluation emphasis
MMSearch (Jiang et al., 2024)	Multimodal search; 300 curated tasks	Human-curated search questions	Requery, rerank, summarize; end-to-end accuracy
Dyn-VQA (Li et al., 2024)	Bilingual VQA; scheduled answer updates	Visual co-reference rewriting; human checks	Adaptive retrieval; multi-hop reasoning
LiveVQA (Fu et al., 2025)	Recent news, videos, and papers	Image relevance; source-context answer filtering	No-search/search comparison; knowledge updating
MMSearch-Plus (Tao et al., 2026)	Multimodal browsing; 311 curated tasks	Lite subset removes items solved without search by any tested model	Provenance; crop/search tools; error analysis
LiveNewsBench (Zhang et al., 2026)	Textual news QA; regular renewal	Source-cluster self-consistency; automated and human checks	No-search/search and budget comparisons
VDR-Bench (Zeng et al., 2026)	Visual multi-hop QA; 2,000 curated tasks	Recorded-path solvability; human checks for textual and prior-based shortcuts	Answer accuracy; intermediate-entity recall
LIVESEARCHVQA (ours)	News VQA; 6,000 items in 30 dated builds	**Versioned sources and joint $V/B/O$ admission records**; repeated panel trials	**Paired CB/WS/OR and trace accounting**; controlled evidence probes