Evaluating Reasoning Under Multimodal Conflict: Separating Score Gains from Evidence Selection

Anonymous authors
Paper under double-blind review

Abstract

Reasoning is used to help multimodal large language models answer questions when visual and textual evidence conflict. Yet higher benchmark scores do not establish more reliable evidence selection, because scores also depend on how answers are elicited and judged. We develop a controlled evaluation protocol that examines correct-source direction, candidate access, and answer scoring through complementary paired comparisons. Its core design crosses reasoning on/off, repeated versus disagreeing incorrect statements, and candidate visibility while holding the required response and output limit fixed within each model. The core comparison covers three open-weight model families, with separate evidence-direction and visual-recognition controls on two. We find that reasoning can improve answers when images are correct while harming them when text is correct. On image-associated knowledge questions, candidates enlarge the reasoning-by-disagreement interaction mainly through a weaker non-reasoning baseline in Qwen, but more through reasoning-enabled accuracy changes in Gemma. The large positive candidate-induced increases in this interaction observed in these two models are not reproduced in the third or under the chart-reading protocol. In the recognition control, the original strict-JSON tests fail, while supplementary wrapper-tolerant scoring reveals label-list benefits and no additional accuracy gain from constrained decoding. These results distinguish format compliance from label matching without attributing vocabulary gains to naming alone. Our findings support evaluating reasoning through source-resolved outcomes, absolute paired accuracies, and explicit answer-information and scoring controls, rather than aggregate gains alone.

1 Introduction

Multimodal large language models (MLLMs) answer questions using images together with text. These sources can disagree: a passage may give the wrong value for a chart, or an image may depict a different entity from the one described. Conflict benchmarks test whether models resist such misleading evidence (Zhang et al., 2025b; Jia et al., 2025; Catapang, 2026). Reasoning can improve benchmark performance, but answer options may also make the task easier to solve (Raman et al., 2026). For conflict evaluation, the important question is what a measured reasoning gain tells us about choosing the correct evidence.

A gain depends on both answers being compared. The reasoning gain is the difference between accuracy with and without reasoning under a given test. Showing candidates can change either accuracy. A larger gain could therefore arise because reasoning produces better answers, because the non-reasoning baseline becomes worse, or both. Averaging across conflict directions creates a second ambiguity: improvement when the image is correct can conceal harm when the text is correct. Our central claim is that these alternatives must be checked before interpreting a score increase as improved evidence selection.

We organize the comparisons around three evaluation choices (Figure 1). **Evidence direction** specifies which source is correct, following the paired conditions in CMC-Bench (Catapang, 2026). The **answer interface** specifies which candidates the model sees and how it must answer; earlier work establishes the importance of options and format (Raman et al., 2026; Rosenthal et al., 2025). **Answer vocabulary** specifies which accepted labels are exposed before generation. These choices overlap:

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.

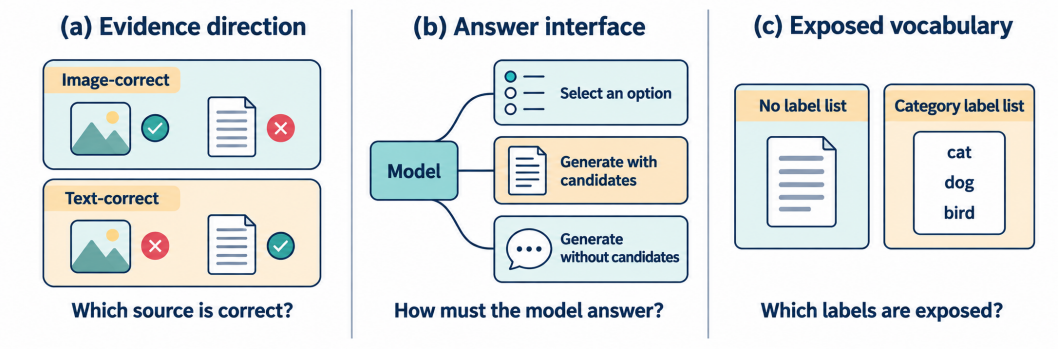


Figure 1: Three factors organizing conflict evaluation. Evidence direction specifies which source is correct under the benchmark construction (checks and crosses), not which source the model uses. The answer interface specifies the candidates shown and response required; exposed vocabulary specifies which labels the model receives. These overlapping factors are studied in separate comparisons, not a full factorial design or stages of a causal process. Inputs and labels are illustrative.

supplying candidates also supplies possible answer names. We do not propose three independent mechanisms or a new conflict benchmark. We use them to separate interpretations that aggregate accuracy leaves unresolved.

Our controlled comparison holds the question, evidence, year-generation instruction, and output limit fixed within each model, and changes only candidate visibility. It crosses this change with reasoning enabled or disabled and with repeated or disagreeing wrong statements. This design asks whether candidates change the additional reasoning gain associated with disagreement. It extends the option-access question studied by Raman et al. (2026) and Liu et al. (2026a) to a controlled wrong-evidence comparison, rather than treating an option–generation accuracy gap as new.

The results expose different reasons for an apparently large gain. In Qwen3.8-27B, candidates mainly enlarge the temporal interaction by lowering the baseline without reasoning. In Gemma-4-31B-it, improved reasoning-enabled answers contribute more. The smaller GLM-4.6V-Flash does not reproduce the positive candidate contrast. On a separate chart cohort, all three estimates are small and their intervals include zero. We also compare two ways of scoring the same pet-recognition outputs: strict JSON compliance and name matching after removing one complete Markdown wrapper. The resulting contrast separates formatting failures from incorrect names, while retaining the extra answer information supplied by a class list.

Our contributions are threefold.

1. We define a conflict-evaluation procedure that reports correct-source direction, absolute paired accuracies, candidate access, and answer acceptance separately.
2. We show how large accuracy-scale reasoning interactions can arise from different baseline changes, using controlled temporal comparisons, a third-model extension, and a chart-task boundary test.
3. We provide a label-list control with both strict and supplementary wrapper-tolerant scores, illustrating why format compliance and answer matching should not be interpreted as the same capability.

The evidence supports these measurement distinctions in the tested settings. It does not identify a shared internal mechanism: model size, native reasoning, task content, and scoring differ across the complementary studies.

2 Related Work

Multimodal conflict and source preference. HallusionBench separates language hallucination and visual illusion (Guan et al., 2024); MMMC studies robustness to modality conflict and training-based improvements (Zhang et al., 2025b). Work on textual misinformation (Zhang et al., 2026) and rela-

tive unimodal uncertainty (Zhang et al., 2025a) examines when models follow one modality. Source preference is not itself evidence of reliability: the preferred source may be wrong. CMC-Bench distinguishes aligned, image-correct, text-correct, and both-wrong conditions (Catapang, 2026), which we use to resolve reasoning effects by direction. MMKC-Bench already compares multiple-choice and open answers under context–memory and inter-context conflict (Jia et al., 2025); we inherit its tasks and options, not a new benchmark. CAER develops conflict-aware routing (Liu et al., 2026b); we do not reproduce it or claim a performance advantage.

Reasoning with answer options. Raman et al. (2026) show that reasoning after seeing options can exploit their information and outperform free-text answering. ReVeL studies verifiable open questions and per-option verification, which retains option content, alongside option-related score inflation (Liu et al., 2026a). We do not claim the option–generation gap or reasoning–option dependence as new. Our narrower comparison crosses candidate visibility with reasoning and repeated versus disagreeing wrong evidence, keeping the generated answer format and within-model budget fixed. Its four-cell accuracies reveal whether an enlarged accuracy-scale interaction comes from a weaker off baseline or a stronger medium contrast. Appendix E compares these designs; neither this distinction nor a large interaction establishes a new general theory of reasoning.

Formatting and scoring. FormatSpread establishes sensitivity to meaning-preserving prompt changes (Sclar et al., 2024). Rosenthal et al. (2025) find that multiple-choice VQA rankings remain sensitive to format even after controlling option order. Hua et al. (2025) show that evaluation choices can contribute to apparent prompt sensitivity. These studies motivate separating supplied answer information from formatting and scoring. We retain exact-match results, show parsing and abstention sensitivities, and report absolute accuracies as well as reasoning effects; we do not infer latent reliability from a single scoring rule.

Recognition and answer vocabulary. Encoded visual information need not be faithfully expressed in generated classifications (Zhang et al., 2024); internal probes and activation interventions also examine visual–linguistic conflict (Nooralahzadeh et al., 2026). Our behavioral controls do not identify an equivalent internal mechanism. Supplying source values, oracle candidates, or a category list changes the available information. The label-list experiment therefore tests an evaluation aid, not a new recognition method, and cannot attribute its benefit solely to naming.

3 Evaluation methodology

Problem and outcomes. Given a question, image, and accompanying statements, we evaluate whether the model's final answer matches the benchmark reference. Evidence selection is the intended capability; reference accuracy is an observable outcome, not a direct measurement of which source the model trusted. We distinguish a gain from enabling reasoning within one interface from a change in that gain across interfaces. CMC uses exact years, otherwise numeric tolerance or normalized text; temporal questions use exact parsed years; ChartQA uses 5% numeric tolerance; pets use normalized label equality. We report these outcomes separately rather than pool them into a reliability score.

Models and coverage. The original comparisons use unmodified Qwen3.8-27B and Gemma-4-31B-it, two distinct model families with native reasoning controls (Qwen Team, 2026; Gemma Team, 2026). A subsequent extension uses the smaller, 9B-class GLM-4.6V-Flash (Z.ai, 2025) on the temporal and chart core only. It adds a model family, not a size-matched comparison; CMC and pets still use two models. We refer to reasoning enabled as *on* (the original records call it *medium*) and disabled as *off*. These native modes need not implement the same algorithm. No comparison changes model weights.

Evidence direction. CMC-Bench supplies 842 paired source examples per model, excluding 100 previously used development examples: 450 from ChartQA and 392 from MMMU (Masry et al., 2022; Yue et al., 2024). Aligned, Image-correct, and Text-correct conditions measure answer accuracy. Both-wrong instead measures rule-detected abstention, excluding empty and truncated outputs. We compare on-minus-off outcomes separately for each condition (Appendix A).

Controlled candidate visibility. We use all 223 temporal/time tasks in the MMKC manifest (149 temporal, 74 time; 219 source groups), without selecting on model correctness. *Duplicate* repeats one constructed incorrect statement; *Mixed* replaces the second with another incorrect answer. Neither

supplies a correct statement. The question image remains, but these image-associated knowledge questions do not require reading the year from the image. The core crosses hidden/visible candidates, off/on reasoning, and Duplicate/Mixed evidence: eight cells, each with 223 tasks. Hidden and visible prompts differ only by an inserted four-option block; both require a four-digit year. Benchmark candidates include the reference, not answers obtained by model retrieval.

Qwen and Gemma additionally cross visibility with a generic four-digit decoding constraint and include a multiple-choice control, for 20 cells and 4,460 requests per model. GLM runs only the eight-cell core (1,784 requests). Caps are fixed within model: 6,144 output tokens for Qwen and 8,192 for Gemma and GLM, including reasoning and final answers. Temperature is zero, with one generation per condition. These are not equal-compute or common-limit comparisons across models. Figures 6–8 illustrate the designs; Appendix B gives full prompts and runtime settings.

Interaction versus final accuracy. Let $A_{m,r,c}$ denote accuracy for reasoning mode m, candidate interface r, and evidence condition c. Define

$$\Delta_{m,r} = A_{m,r,\text{Mixed}} - A_{m,r,\text{Duplicate}}, \qquad \Gamma_r = \Delta_{\text{on},r} - \Delta_{\text{off},r}, \qquad D = \Gamma_{\text{visible}} - \Gamma_{\text{hidden}}.$$

Γ_r measures how much more reasoning helps under Mixed than Duplicate evidence. D asks whether candidate visibility changes that contrast. A positive D is not a D-point increase in final accuracy. We therefore report all eight accuracies and separate the on and off contributions.

Chart and recognition controls. The chart study uses 256 images (128 human-authored and 128 machine-generated questions), one numeric question per chart, selected without model-output screening (Masry et al., 2022). The prompt prioritizes the chart. Qwen and Gemma have the eight wrong-evidence core cells plus four repeated-correct-statement controls; GLM has the core only. All retain their temporal caps. GLM's chart parser additionally accepts a fixed box wrapper, specified before its complete run after a gold-blind format check; it is not parser-identical to the original two-model comparison. The separate Oxford-IIIT Pet control uses 185 test images, five per breed, and a 37-label trainval vocabulary (Parkhi et al., 2012). Its plain, list-only, and list-plus-constrained-decoding arms disable reasoning and use 256 tokens. Hash and near-duplicate checks exclude project-history image overlap, not pretraining exposure. These are complementary cohorts, not a controlled intervention on task type.

Failures and uncertainty. Every cell retains its full denominator; capped, empty, or unparseable final answers score zero, with no answers extracted from reasoning. Execution integrity and answer validity are separate: a complete run may contain unsuccessful answers. The original parsing-threshold failures remain recorded, including GLM's failed temporal admission check. GLM was subsequently collected and scored as an end-to-end extension, not declared to have passed that check.

Paired bootstrap intervals use 10,000 draws of source groups for temporal tasks, images for ChartQA, and images within 37 fixed pet classes. They measure sample uncertainty, not repeated-run variance. We retain the original two-model 97.5% intervals and criteria; GLM's later intervals are descriptive, not a retrospective third confirmatory test. The pet wrapper rule is post-hoc for Qwen and was fixed before Gemma's collection, without replacing either strict-JSON endpoint. Appendices B–C document revisions, invalid-answer bounds, and the timing of each analysis.

4 Reasoning can improve image-correct answers while harming text-correct answers

Reasoning does not help equally across the conflict conditions we test. CMC lets us ask whether it improves answers when either the image or the text is the correct source, rather than only when one modality is misleading.

Which source is correct matters. Figure 2 reports CMC results, where the image can supply the answer. Medium reasoning improves Gemma's image-correct accuracy by 2.61 points and its text-correct accuracy by 19.83 points. Qwen improves image-correct accuracy by 11.88 points but reduces text-correct accuracy by 7.13 points and aligned accuracy by 7.96 points. When both sources are wrong, rule-detected abstention excluding empty and truncated outputs falls from 9/842 to 0/842 on Gemma and stays at 2/842 on Qwen (Appendix A). The direction-resolved changes are consistent

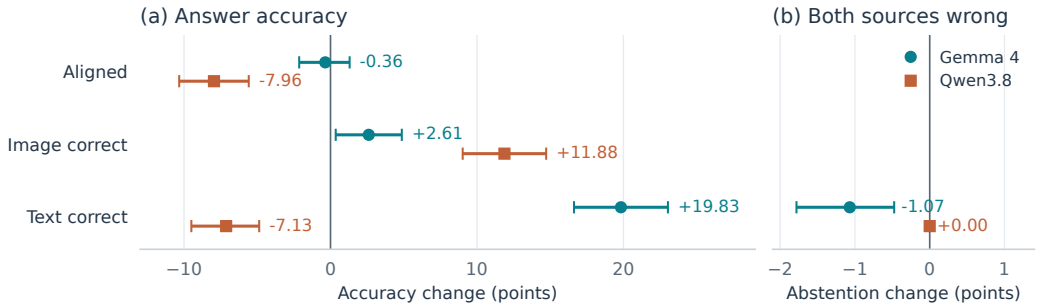


Figure 2: Reasoning on minus off on CMC (842 source examples per model). (a) Answer accuracy, separated by which source is correct. Text-correct accuracy improves on Gemma and worsens on Qwen. (b) Rule-detected, nonempty, non-truncated abstention when both sources are wrong; note the separate horizontal scale. Reference matching is not an abstention metric (Appendix A). Whiskers are source-example bootstrap 95% intervals.

Table 1: CMC absolute outcomes on 842 source examples. Off and Medium are percentages; changes are points with paired 95% intervals. The first three rows per model measure answer accuracy; Both wrong measures valid rule-detected abstention, not reference matching.

Model	Evidence condition	Off	Medium	Change [95% CI]
Qwen3.8	Aligned	89.07	81.12	-7.96 [-10.33, -5.58]
	Image correct	31.24	43.11	+11.88 [9.03, 14.73]
	Text correct	87.53	80.40	-7.13 [-9.50, -4.87]
	Both wrong (abstain)	0.24	0.24	+0.00 [0.00, 0.00]
Gemma 4	Aligned	78.74	78.38	-0.36 [-2.14, 1.31]
	Image correct	33.25	35.87	+2.61 [0.36, 4.87]
	Text correct	63.78	83.61	+19.83 [16.63, 23.04]
	Both wrong (abstain)	1.07	0.00	-1.07 [-1.78, -0.48]

with different shifts in source preference, but do not establish which incorrect source a response follows. A positive mean conflict score is therefore insufficient evidence of uniformly more reliable source selection.

The paired answers make the harm concrete. Among text-correct CMC examples, Gemma corrects 196 errors but changes 29 correct answers to incorrect ones; Qwen corrects only 21 and makes 81 previously correct answers wrong. On aligned examples, Qwen similarly corrects 23 errors but introduces 90. These counts do not tell us whether every newly incorrect answer follows the image: the released records omit the answer associated with the substituted image.

These results concern evidence direction under a fixed answer interface. A separate question is whether the interface itself changes the measured benefit of reasoning. We next test candidate visibility while holding the required response fixed.

5 Candidates can enlarge reasoning gains by weakening the baseline

The positive candidate contrast is large on Qwen and Gemma, but not reproduced on GLM. On the 223 temporal tasks, D is 45.29 points for Qwen (95% interval [36.77, 54.02]) and 17.04 for Gemma [10.82, 23.25]. Their original adjusted 97.5% intervals are [35.45, 55.20] and [9.91, 24.23]. GLM's later estimate is −5.83 points [−13.72, 2.26], which does not establish a negative population effect. The comparison concerns adding this candidate block within each model, not pure output format or model capability rankings.

The baseline explains much of Qwen's large interaction. Under Mixed evidence, Qwen's off accuracy falls from 56.50% without candidates to 18.39% with them, while on accuracy rises from 73.54% to 82.51% (Figure 3). After accounting for Duplicate, 36.32 of the 45.29 points come from

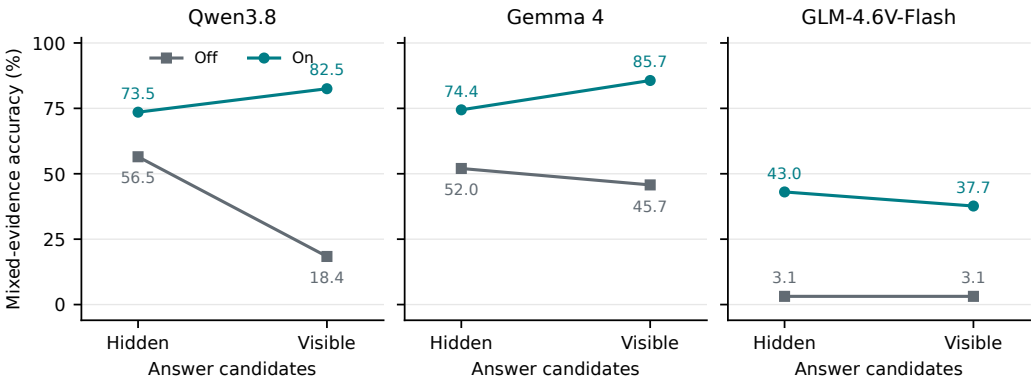


Figure 3: Accuracy on the same 223 temporal tasks with disagreeing statements (Mixed). Showing candidates lowers off accuracy and raises on accuracy for Qwen and Gemma; GLM does not follow that pattern. A larger on–off gap can therefore reflect a weaker baseline as well as better on-mode answers. Duplicate-condition accuracies and the full contrast appear in Table 2; lines connect paired settings, not intermediate interventions. Output caps are fixed within each model (Qwen 6,144; Gemma and GLM 8,192), not equalized computational budgets.

Table 2: Temporal accuracies (223 tasks per cell). Accuracies are percentages; Γ_r is in points. On denotes each model's native reasoning mode. Qwen uses 6,144 tokens; Gemma and GLM use 8,192. All models generate years, with no answer-specific decoding constraint.

		Off		On		
Model	Candidates	Duplicate	Mixed	Duplicate	Mixed	Γ_r
Qwen3.8	Hidden / generate	2.69	56.50	0.00	73.54	19.73
	Visible / generate	0.90	18.39	0.00	82.51	65.02
Gemma 4	Hidden / generate	0.90	52.02	0.00	74.44	23.32
	Visible / generate	0.45	45.74	0.00	85.65	40.36
GLM-4.6V-Flash	Hidden / generate	0.90	3.14	0.00	43.05	40.81
	Visible / generate	0.90	3.14	0.45	37.67	34.98

the lower off contrast and 8.97 from the higher on contrast. Gemma differs: 11.21 of its 17.04 points come from the on contrast and 5.83 from the off contrast. These are arithmetic components, not causal mediators. The same sign of D can summarize substantially different changes in model behavior.

Options are not uniformly helpful, even when they contain the reference. Mixed candidate effects are −38.12 points on Qwen off and +8.97 on Qwen on; Gemma's are −6.28 and +11.21 (intervals in Appendix B). Repeated wrong statements remain difficult: on accuracy is zero in both ordinary-generation interfaces for these models (Table 2). Thus a large interaction does not imply consistently overcoming misleading evidence; Appendix D gives saved-output examples.

GLM supplies a model boundary, not evidence that reasoning is useless. Its Mixed accuracy rises from 3.14% off to 43.05% on without candidates, and from 3.14% to 37.67% with them. Its reasoning-by-disagreement interactions remain positive (40.81 and 34.98 points); the unreplicated quantity is the positive candidate increment D. Twenty-five of 1,784 outputs are capped without a final answer. A complete 9,500-token run preserves all final answers and failures, not establishing an unlimited-budget ceiling. Assigning arbitrary correctness to invalid answers gives sample-level D bounds [−14.80, −3.59]. These logical bounds are not confidence intervals; the sampling interval still includes zero.

The two positive results survive the tested syntax and failure controls, but not every scale. A generic four-digit constraint leaves visibility contrasts of 42.60 points on Qwen and 17.04 on Gemma. Full-run failures are 12 capped plus one other invalid answer on Qwen, and 26 capped plus five other invalid answers on Gemma. Signed invalid-answer bounds remain [42.15, 45.29] and [10.76, 18.83].

Table 3: ChartQA wrong-evidence accuracies (256 images per cell). Accuracies are percentages; Γ_r is in points. On denotes each model's native reasoning mode. Qwen uses 6,144 tokens; Gemma and GLM use 8,192. GLM accepts its fixed box wrapper; parsing is not identical across models.

		Off		On		
Model	Candidates	Duplicate	Mixed	Duplicate	Mixed	Γ_r
Qwen3.8	Hidden	92.19	91.80	96.09	96.09	0.39
	Visible	95.31	94.14	98.05	98.44	1.56
Gemma 4	Hidden	84.77	85.16	93.75	93.75	-0.39
	Visible	95.31	94.53	97.27	96.88	0.39
GLM-4.6V-Flash	Hidden	86.33	81.64	93.75	92.58	3.52
	Visible	86.33	85.55	97.27	97.27	0.78

These observations limit format-only explanations; they do not remove every consequence of adding tokens to the prompt. On smoothed marginal log-odds, both primary intervals contain zero for all three tested pseudocounts. Thus the result is an accuracy-scale interaction, not evidence of a scale-independent mechanism (Appendix B).

Candidate content matters beyond visibility. An exploratory follow-up on 64 previously examined tasks crosses incorrect statement years with candidate-year lists, keeping the correct year visible. Under Mixed evidence, removing wrong-year overlap raises Qwen's off accuracy from 53.13% to 81.25% and on accuracy from 83.59% to 89.06%, averaged over two statement versions. The reasoning gain therefore shrinks from 30.47 to 7.81 points despite higher final accuracy. Gemma does not show the same clear pattern. This content comparison is distinct from D; it balances additive year-identity effects but is not independent confirmation or a deployable correction (Appendix B.6).

The large positive contrasts do not replicate under the chart protocol. Qwen's D is 1.17 points (original 97.5% interval [−1.56, 3.91]) and Gemma's is 0.78 [−3.52, 5.08]; neither meets the positive replication criterion. GLM's extension gives −2.73 points (descriptive 95% interval [−7.03, 1.17]). Figure 5 places these estimates beside the temporal results without pooling tasks or testing a causal task difference. Unlike temporal Duplicate, chart Duplicate remains highly accurate (Table 3). Chart content, source-prioritizing instructions, distractors, and baseline headroom all change together.

Answer failures still limit content-level interpretation. Gemma has 108 invalid chart outputs among 3,072 requests, including 20 capped; its logical D bounds are [−13.67, 15.23]. A post-hoc numeric-format rule admits 19 more responses without changing the 0.78 estimate, but the remaining bounds [−12.11, 13.67] still permit a large content effect. Qwen has two invalid outputs, including one capped, with bounds [1.17, 1.95]. GLM has six invalid outputs among 2,048 requests, including one capped, and bounds [−3.91, −1.56]; its box-wrapper rule differs from the original parser. Small point estimates are not proof of equivalence, and GLM's two intervals crossing zero do not establish a temporal–chart difference. Full outcomes and parser chronology appear in Appendix C.

6 Label-list gains depend on separating answer content from format

Can a task's label list improve exact match without changing model weights, and how much does the score depend on output formatting? We compare label exposure and decoding constraints on new pet images. This is an image-only recognition control, not a replication of conflict arbitration.

A fixed list on new images. The 37 breed names come from trainval metadata, not the test images or model predictions, and the same alphabetically sorted list is used for all 185 test images. The plain arm has no list. The list arm adds the names; the guided arm uses identical messages but restricts decoding to a JSON-schema enumeration of those names. All arms request a JSON object containing one answer. Reference coverage is complete by task construction: the list supplies possible answers and may assist recognition, search, and naming together.

Appendix D shows original outputs for a label-list repair and a regression, each scored under both rules.

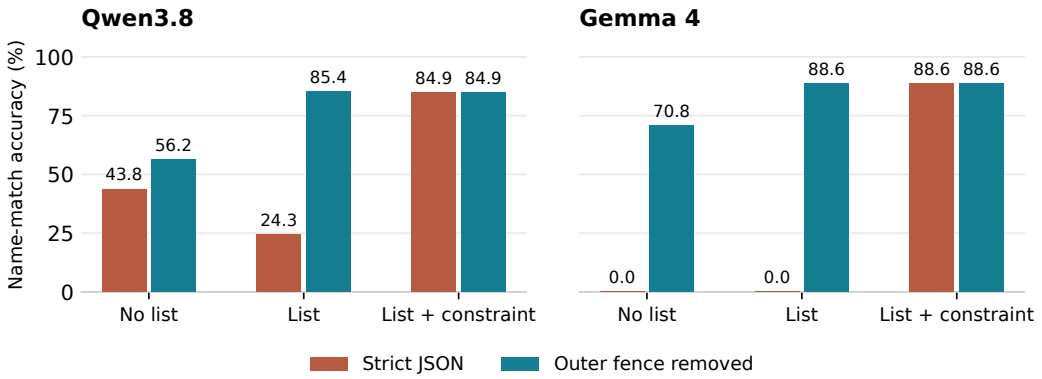


Figure 4: The same pet-recognition outputs under two parsers (185 images per arm). Strict JSON scoring confounds name matching with Markdown wrapping. Removing one complete outer fence reveals list benefits but no positive constrained-decoding increment. This supplementary rule is post-hoc for Qwen and fixed before Gemma's collection; it does not replace either original primary test. Paired intervals and counts are reported in Appendix C.

Strict JSON and name matching give different conclusions. The frozen strict parser accepts only an unwrapped JSON object. Under this rule, plain/list/guided accuracies are 43.78/24.32/84.86% on Qwen and 0/0/88.65% on Gemma. Neither model passes the original positive list-benefit criterion. All 370 ordinary-generation Gemma answers, and 170 Qwen answers, are complete JSON objects enclosed in Markdown fences. These failures reflect the requested format, not necessarily incorrect names.

A supplementary rule removes only one complete outer JSON or unlabelled Markdown fence, then applies the unchanged JSON parser and symmetric NFKC, whitespace, and case normalization. It does not extract prose, repair JSON, or add aliases. This rule was chosen after Qwen's results and sealed before Gemma's first output; it is not the original primary endpoint for either model. The adjusted accuracies are 56.22/85.41/84.86% for Qwen and 70.81/88.65/88.65% for Gemma (Figure 4). List-minus-plain gains are 29.19 points (97.5% descriptive interval [22.16, 35.68]) and 17.84 [12.43, 23.24]. Qwen's 21 remaining empty objects still count as failures.

Decoding constraints mainly change format compliance in this comparison. Once wrappers are accepted, guided-minus-list differences are −0.54 points on Qwen and zero on Gemma, rather than the large gains under strict JSON scoring. The zero observed difference is not population equivalence. The list also harms some answers: 12 Qwen and five Gemma images change from correct to wrong despite the positive net gains. Appendix C reports counts, paired transitions, and both scoring rules. This supports distinguishing name matching from format compliance, not a pure naming mechanism or a claim that constrained decoding is generally unnecessary.

7 Discussion and Conclusion

A reasoning gain describes a change from a baseline. It does not reveal, by itself, whether the model has become better at choosing evidence. Our comparisons show three concrete reasons. Averaging CMC directions conceals harm to Qwen's text-correct answers. In temporal questions, candidates enlarge the interaction mainly by weakening Qwen's off baseline, whereas Gemma benefits more from its on contrast; GLM does not reproduce the positive candidate increment D. In pet recognition, accepting complete wrappers changes which apparent gains concern names and which concern formatting.

From findings to an evaluation report. Table 4 lists four checks for interpreting gains, not a new reliability score.

For a new model, compare off and on with the same interface, scoring rule, and output cap. Vary candidate access without changing the required response, and inspect absolute accuracies.

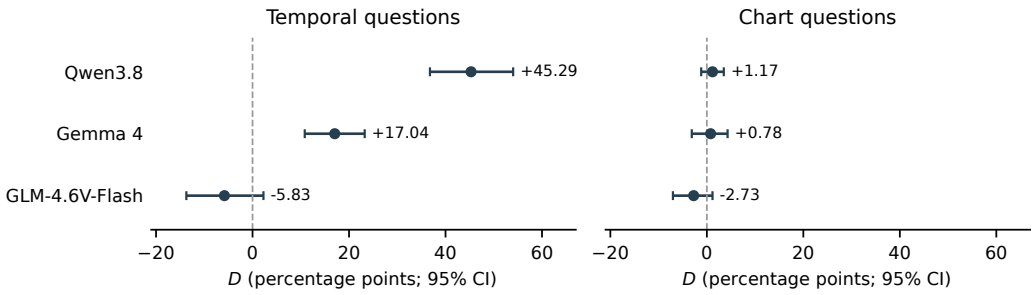


Figure 5: Candidate-visibility contrast D across models and tasks, in percentage points. Bars are paired 95% bootstrap intervals (219 source groups for temporal tasks; 256 images for charts). Qwen and Gemma have positive temporal contrasts; GLM's interval includes zero. All three chart intervals include zero. These are not equivalence tests or causal tests of task differences. The original two-model decisions use 97.5% intervals; the later GLM estimates are descriptive and use a model-specific chart wrapper rule (Appendix B.5).

Table 4: Reporting questions motivated by the observed failure modes. These checks qualify the interpretation of a gain; they do not certify reliable evidence selection.

Question	What to report	Why it matters here
Which source benefits?	Paired off/on accuracy by correct source, aligned retention, and valid abstention when both sources are wrong.	A positive average conceals directional harm (Figure 2).
Where does the gain come from?	Absolute accuracies in all eight cells, with paired contrasts; candidate origin and reference coverage.	A larger interaction can reflect a weaker off baseline (Table 2).
Content or format?	Format compliance and answer matching separately; parser rules, their timing, and failures in the full denominator.	Wrappers alter apparent label and decoding gains (Figure 4).
How far does it extend?	Model- and task-specific intervals, output caps, invalid-answer bounds, and unsuccessful replications.	Large positive contrasts are not reproduced in every tested setting (Figure 5).

Report scoring-dependent conclusions alongside the original endpoint. These checks do not certify evidence selection: correct answers may come from prior knowledge or candidates rather than the intended source. Stronger claims require interventions that distinguish those routes.

Scope. The temporal tasks are synthetic, historically examined knowledge questions, not a test of reading years from pixels. The two positive candidate increments are accuracy-scale results; log-odds sensitivity does not establish the same interaction. The chart estimates do not reproduce these increments, but high baselines and unresolved Gemma outputs constrain the interpretation. Task content and prompts change together, so we cannot identify why the results differ across tasks.

Three model families do not represent the population of MLLMs: GLM is smaller, native reasoning and caps differ, and each condition has one generation. GLM's chart wrapper amendment also prevents a parser-identical three-model comparison. CMC and pets still cover only Qwen and Gemma. Pet labels are not independently adjudicated, and the supplementary wrapper rule was informed by Qwen's outputs. Image-overlap checks do not exclude pretraining exposure.

Conclusion. Reliable evidence selection requires more than a positive aggregate reasoning gain. The relevant evidence is in the source-resolved outcomes, the absolute accuracies behind the gain, and the information and formatting rules supplied by the test. Our results make these distinctions measurable while showing where the observed effects fail to generalize.

AI USE STATEMENT

AI coding and writing assistants contributed to experiment engineering, job monitoring, literature discovery, interpretation drafts, prose editing, figure-generation code, and simulated reviewer feedback. Simulated reviews are internal editorial checks, not independent human peer review. Experiment records retain protocols, outputs, and dated revisions. Quantitative results are computed by the recorded evaluators; an assistant's interpretation is not a substitute for those records or for independent validation.

The quantitative figures are programmatically plotted from saved numerical artifacts. The conceptual overview in Figure 1 and the experimental-design schematics in Figures 6–8 were generated with an image model and checked for label accuracy and consistency with the recorded protocols. They contain no experimental measurements and do not represent benchmark images or model outputs. Machine-assisted semantic or label audits are explicitly diagnostic where independent human adjudication has not been performed. The human authors retain responsibility for checking the results, references, originality, disclosures, and submission compliance. This draft does not assert that every assistant-proposed decision has already received individual human verification.

ETHICS STATEMENT

The work uses public benchmark releases and synthetic evidence interventions, with no recruited human participants or deployment study. Public availability does not by itself guarantee that images contain no personal information; redistribution must respect the source licenses and privacy constraints. Machine-assisted label diagnostics are not independently adjudicated human annotations. The controlled benchmark comparisons do not establish safe real-world evidence arbitration. We preserve negative results and distinguish task-protocol changes from model improvements.

REPRODUCIBILITY STATEMENT

The original two-model primary protocols were fixed before their corresponding results were examined; outputs were saved and hashed after evaluation. The later GLM extension, follow-ups on previously examined examples, and post-hoc analyses are labelled separately. The project retains prompts, configurations, raw outputs, analysis scripts, and file hashes; a local results-level audit candidate has been prepared, but a complete anonymized inference-reproduction archive and its public release remain submission preparation tasks. Quantitative plots use named fields from frozen results or separately recorded post-hoc scoring audits, with source hashes; Appendix A identifies the corrected abstention metric. Appendices B–C document the complete current comparisons, including failed replications, invalid answers, and scoring amendments; Appendix B.6 separates the exploratory candidate-content controls from the core comparison. Appendix F describes the retained materials and remaining release work.

REFERENCES

Jasper Kyle Catapang. When image and text disagree: Cross-modal evidence conflict in multimodal retrieval-augmented generation. In *Proceedings of the 2nd Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2026)*, pp. 1–10, 2026. doi: 10.18653/v1/2026.magmar-main.3. URL https://aclanthology.org/2026.magmar-main.3/.

Gemma Team. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026. URL https://arxiv.org/abs/2607.02770.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html.

Andong Hua, Kenan Tang, Chenhe Gu, Jindong Gu, Eric Wong, and Yao Qin. Flaw or artifact? rethinking prompt sensitivity in evaluating LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19889–19899, 2025. doi: 10.18653/v1/2025.emnlp-main.1006. URL https://aclanthology.org/2025.emnlp-main.1006/.

Yifan Jia, Kailin Jiang, Yuyang Liang, Qihan Ren, Yi Xin, Rui Yang, Fenze Feng, Mingcai Chen, Hengyang Lu, Haozhe Wang, Xiaoye Qu, Dongrui Liu, Lizhen Cui, and Yuntao Du. Benchmarking multimodal knowledge conflict for large multimodal models. *arXiv preprint arXiv:2505.19509*, 2025. URL https://arxiv.org/abs/2505.19509.

Yesheng Liu, Hao Li, Haiyu Xu, Baoqi Pei, Jiahao Wang, Mingxuan Zhao, Jing-Shu Zheng, Zheqi He, JG Yao, Bowen Qin, Xi Yang, and Jiajun Zhang. Beyond multiple choice: Verifiable OpenQA for robust vision-language RFT. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a. URL https://arxiv.org/abs/2511.17405v2. Design comparison uses arXiv:2511.17405v2.

Zixuan Liu, Juntao Cai, Xiaoxu Cai, Haishuai Wang, and Jiajun Bu. CAER: Conflict-aware evidence routing with dual prefix experts for multimodal large language models. *arXiv preprint arXiv:2607.28991*, 2026b. URL https://arxiv.org/abs/2607.28991.

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

Farhad Nooralahzadeh, Omid Rohanian, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. Arbitration failure, not perceptual blindness: How vision-language models resolve visual-linguistic conflicts. *arXiv preprint arXiv:2604.09364*, 2026. URL https://arxiv.org/abs/2604.09364.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012. doi: 10.1109/CVPR.2012.6248092. URL https://www.robots.ox.ac.uk/~vgg/publications/2012/parkhi12a/parkhi12a.pdf.

Qwen Team. Qwen3.8-27B. https://huggingface.co/Qwen/Qwen3.8-27B, 2026.

Narun K. Raman, Taylor Lundy, and Kevin Leyton-Brown. Reasoning models are test exploiters: Rethinking multiple-choice. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2507.15337v2. Design comparison uses arXiv:2507.15337v2.

Fabio Rosenthal, Sebastian Schmidt, Thorsten Graf, Thorsten Bagdonat, Stephan Günnemann, and Leo Schwinn. Unexplored flaws in multiple-choice VQA make benchmarking unreliable. *arXiv preprint arXiv:2511.22341*, 2025. URL https://arxiv.org/abs/2511.22341v3. Version 3, revised September 2026; accepted to Findings of EMNLP 2026.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6c0e99d736da621403018ca7b32b1a4d-Paper-Conference.pdf.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

Z.ai. GLM-4.6V-Flash model card. https://huggingface.co/zai-org/GLM-4.6V-Flash, 2025. Official model card, accessed September 23, 2026.

Chi Zhang, Wenxuan Ding, Jiale Liu, Mingrui Wu, Qingyun Wu, and Ray Mooney. Do images speak louder than words? investigating the effect of textual misinformation in VLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6872–6895, 2026. doi: 10.18653/v1/2026.eacl-long.323. URL https://aclanthology.org/2026.eacl-long.323/.

Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5c7024041be305c94d7311cfcc53d93e-Abstract-Conference.html.

Zhuoran Zhang, Tengyue Wang, Xilin Gong, Yang Shi, Haotian Wang, Di Wang, and Lijie Hu. When modalities conflict: How unimodal reasoning uncertainty governs preference dynamics in MLLMs. *arXiv preprint arXiv:2511.02243*, 2025a. URL https://arxiv.org/abs/2511.02243.

Zongmeng Zhang, Wengang Zhou, Jie Zhao, and Houqiang Li. Robust multimodal large language models against modality conflict. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 77233–77253. PMLR, 2025b. URL https://proceedings.mlr.press/v267/zhang25dq.html.

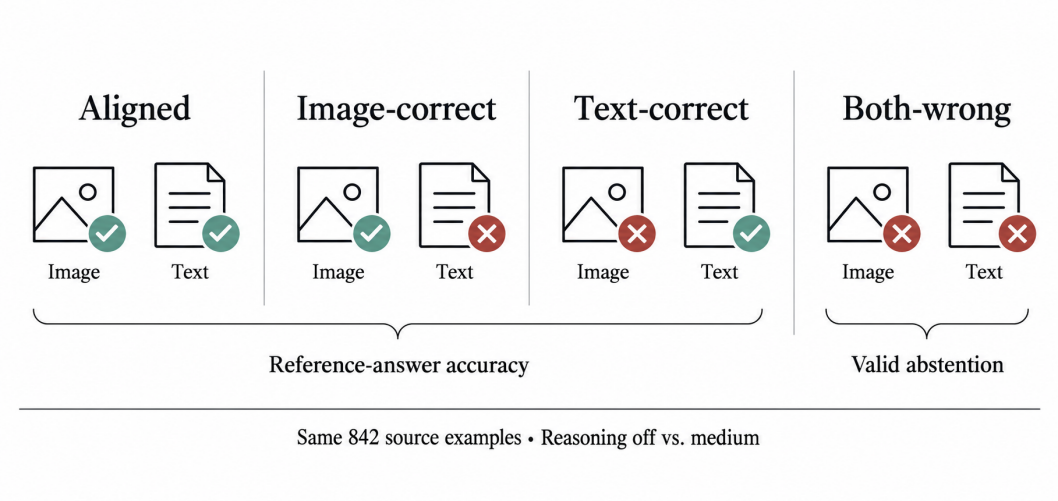


Figure 6: Evidence conditions in the direction-resolved comparison. Checks and crosses denote correct and incorrect sources under the benchmark construction, not observed source use. The Both-wrong endpoint is valid, non-truncated rule-detected abstention; it must not be pooled with reference-answer accuracy in the other three conditions. Icons are illustrative, not benchmark images.

Guide to the supporting evidence. Appendix A documents the direction comparison and its abstention correction; Appendix B gives the temporal controls, sensitivities, and exploratory candidate-content comparisons; Appendix C reports the chart non-replication and both pet-scoring rules. Selected outputs illustrate these results in Appendix D. The final two sections compare the closest prior designs and describe the retained reproduction materials.

A Evidence direction: protocol and scoring

A.1 Cohort and request

The CMC-Bench release contains 942 source examples. Excluding the fixed 100-example development split leaves 842 (450 ChartQA and 392 MMMU), each evaluated under four evidence conditions and two reasoning modes. No answer candidates are shown. The original passage wording can reveal whether text was replaced, so we remove it in every condition. The single user message is:

> Visual evidence:
> *[condition image]*
> Text evidence states that the answer is: *[text answer]*
> Question: *[question]*
> Use the evidence to answer the question. If the sources conflict, decide which source actually answers the question. Return only a short answer in the final response, without explanation.

The condition labels describe benchmark construction, not independently verified source truth or observed model behavior. The released records omit answers associated with substituted images, limiting attribution of an incorrect response to that source.

A.2 Runtime and answer scoring

Both models use BF16 weights, vLLM 0.27.1, a 6,144-token output limit, temperature zero, seed 20260903, and serial requests. Revisions and native off/medium settings are listed in Appendix B; the CMC runtime otherwise differs from that later study. Qwen uses a 262,144-token context, FP8 KV cache, enabled prefix caching, a 2,097,152-pixel image limit, and MTP speculative decoding with one speculative token. This MTP setting is specific to the historical CMC run; its effect is not isolated. Gemma uses a 12,288-token context, BF16 KV cache, default image preprocessing, and

disabled prefix caching. The reported Gemma run was separately fixed after the earlier cache-enabled configuration failed repeatability checks; caching is not proposed as a capability intervention.

Scoring requires exact calendar years, allows 5% relative tolerance for other purely numerical answers, and uses normalized equality for text. This is automatic scoring, not independent semantic adjudication. Table 1 reports all absolute outcomes. Paired intervals resample the 842 source examples jointly across conditions and modes (10,000 replicates, seed 20260903).

A.3 Both-wrong is an abstention endpoint, not reference matching

The original analysis matched the reference in all four conditions. An earlier caption incorrectly described that Both-wrong outcome as abstention. Figure 2 and Table 1 instead use rule-detected, nonempty, non-truncated abstention for Both-wrong; the first three conditions remain answer accuracy. The phrase-based detector is not independent semantic adjudication.

Table 5: Both-wrong scoring reconciliation, 842 examples per model. Counts are successes under each rule; Δ is medium minus off in points with paired 95% intervals. Only abstention supplies the Both-wrong endpoint in the main text.

Model	Scoring rule	Off	Medium	Δ [95% interval]
Gemma	Original-reference match	25	27	0.24 [-0.83, 1.31]
	Valid rule-detected abstention	9	0	-1.07 [-1.78, -0.48]
Qwen	Original-reference match	23	30	0.83 [-0.24, 1.90]
	Valid rule-detected abstention	2	2	0.00 [0.00, 0.00]

Gemma has zero newly valid abstentions and nine lost abstentions; Qwen has neither. The legacy abstention function also accepted seven empty, token-limited Gemma medium outputs, which are failures under the corrected rule. Matching the reference when neither supplied source contains it is reported separately, not credited as abstention.

B Candidate visibility and content: temporal comparisons

The following protocol documents the original Qwen/Gemma controlled temporal study, not a reanalysis or replacement of the historical interface outputs. The later GLM extension is described separately in Appendix B.5. The main candidate-visibility contrast was specified before the two-model runs; the 223 tasks had already appeared in the earlier studies. No task was selected by model correctness. The study is a controlled follow-up on the same task population, not an independent sample replication.

B.1 Request construction and candidate provenance

For each task, the fixed historical source assignment supplies two incorrect statements. Duplicate repeats the base statement; Mixed uses the alternate statement second. The request contains, in order, `External evidence is provided below.`, the two numbered statements, `Question image:`, the question image, and the question text. The ordinary-generation instruction is identical in hidden and visible conditions:

> Answer with only one four-digit year. Do not provide options or an explanation in the final answer.

The visible condition inserts `Options:` followed by four labelled benchmark answer values between the question and that instruction. As in MMKC's construction (Jia et al., 2025), the options contain the reference, the two conflicting values, and another distractor. The fixed per-task option block is shared across all visible conditions. Removing that block reproduces the hidden prompt exactly. Thus visibility changes available answer information and prompt length together; it does not isolate semantic candidate information from all token-level effects. External evidence images are absent in

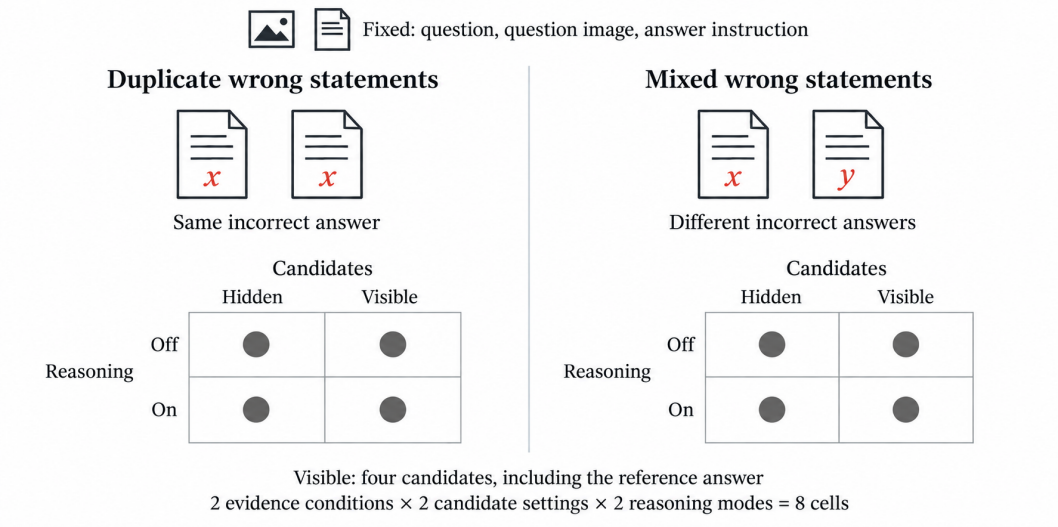


Figure 7: Paired candidate-visibility designs. Each dot marks one experimental condition, not a measured result. The eight-cell comparison crosses evidence composition, candidates, and reasoning; *x* and *y* are distinct incorrect answers. Visible candidates include the reference, while neither text condition does. All three models run this core design on temporal and chart tasks. Qwen and Gemma additionally have syntax/MCQ controls (temporal) and repeated-correct controls (charts). Both studies retain the question image, but only the chart prompt explicitly requires using it; task type is not a controlled intervention. *D* uses only the eight ordinary-generation cells.

both conditions. The retained question image identifies the subject; these knowledge questions do not constitute a controlled test of reading the answer from the image.

For the syntax control, prompt content is unchanged and the backend constrains the final answer with the generic regular expression `[0-9]{4}`. It supplies no task-specific year. The multiple-choice condition retains the option block but replaces the instruction with:

> Return only the option letter A, B, C, or D as the final answer.

Its constrained answer set is A–D. This control jointly changes the response instruction and answer set; it is not a pure letter-encoding intervention.

Appendix D gives complete statements, options, and saved final answers for two selected tasks. Option order is fixed per task; earlier order controls do not establish invariance to every order in this setting.

Weights and KV caches use BF16, with no quantization. Both deployments use one GPU and a 12,288-token context. Qwen uses a fixed 6,144-token output limit and Gemma 8,192, with temperature zero and seed 20260826. Each native reasoning mode is configured separately; off disables explicit thinking. Request hashes reconstruct the prompt, image, decoding settings and model-specific reasoning parameters. All 4,460 study requests per model are distinct (16 factorial and four secondary MCQ conditions per task); repeated engineering probes are excluded from scientific denominators.

Exact mode and runtime settings. The label *medium* is retained from the experiment records; it is not an identical API setting in both models. Qwen sets `enable_thinking=true`, `preserve_thinking=true`, and `reasoning_effort=medium`; off sets `enable_thinking=false`, retains `preserve_thinking=true`, and omits the effort setting. Gemma sets only `enable_thinking=true` or `false`, with no effort parameter in either mode. The reasoning parsers are `qwen3` and `gemma4`. Both runs use vLLM 0.27.1+cu129, PyTorch 2.13.0+cu129, and Transformers 5.15.1, with serial requests, chunked prefill, and prefix caching disabled. Qwen uses a 2,097,152-pixel image-processor limit; Gemma uses its snapshot's default processor. Model revisions are `1d4bf0f2ff6012fd82039f2fa52739d0dd7c60c0` for Qwen and `842da3794eaa0b77d5f08bae87a17459d91ff475` for Gemma.

Table 6: Controlled temporal comparison: all 223 tasks in each cell. Accuracies are percentages; Γ_r is in points. Each model uses one fixed output limit (Qwen: 6,144; Gemma: 8,192). Both evidence conditions contain only wrong statements.

		Off		Medium		
Model	Candidate / response	Duplicate	Mixed	Duplicate	Mixed	Γ_r
Qwen3.8	Hidden / generate	2.69	56.50	0.00	73.54	19.73
	Visible / generate	0.90	18.39	0.00	82.51	65.02
	Hidden / syntax	2.69	56.50	0.00	73.54	19.73
	Visible / syntax	0.90	21.08	0.00	82.51	62.33
	Visible / MCQ	0.45	7.17	0.45	83.86	76.68
Gemma 4	Hidden / generate	0.90	52.02	0.00	74.44	23.32
	Visible / generate	0.45	45.74	0.00	85.65	40.36
	Hidden / syntax	0.90	52.02	0.00	74.44	23.32
	Visible / syntax	0.45	45.74	0.00	85.65	40.36
	Visible / MCQ	0.90	0.90	0.00	83.41	83.41

Table 7: Paired accuracy-scale contrasts (points). The last column varies all invalid answers with the sign of each contrast coefficient; it is not a confidence interval.

Model	Contrast in Γ	Estimate [95% CI]	Invalid-answer bounds
Qwen3.8	Visible − hidden, generate	45.29 [36.77, 54.02]	[42.15, 45.29]
	Visible − hidden, syntax	42.60 [34.38, 51.12]	[39.91, 42.60]
	MCQ − visible generate	11.66 [5.00, 18.18]	[11.66, 12.11]
Gemma 4	Visible − hidden, generate	17.04 [10.82, 23.25]	[10.76, 18.83]
	Visible − hidden, syntax	17.04 [10.82, 23.25]	[12.11, 17.94]
	MCQ − visible generate	43.05 [35.91, 50.22]	[42.15, 44.39]

Only the final-answer stream is scored. The frozen year parser accepts a four-digit year beginning with 1 or 2, optional answer prefixes and punctuation, or a single unique such year in the final text; multiple distinct years are unparseable. It never extracts answers from the reasoning stream. A token-capped answer is incorrect even if a year can be extracted.

B.2 Amendment and estimand

Earlier Gemma runs stopped because a small number of outputs exhausted the limit without a parseable final answer. The larger-limit version also failed the original parsing threshold. Before the new full run and any controlled temporal study scientific scoring, the protocol was amended to measure success under a fixed output budget: truncated, empty, or unparseable final answers score zero. These outputs are not dropped, repaired, or replaced. The original failed runs are kept separate, and the amended full run starts from the beginning. Gemma still fails the legacy parsing criterion; passing completeness and execution checks is not a retrospective pass of that criterion.

Qwen's completed 6,144-token run uses the same all-denominator scoring convention. A common-limit Qwen rerun was not conducted, so the two model effects must not be interpreted as a matched-budget comparison. All contrasts concern the observed limits, not unlimited reasoning or equal computational cost.

B.3 Inference and failure sensitivity

Each bootstrap replicate samples 219 source groups with replacement, retaining all tasks and conditions in each sampled group. Accuracy is task-weighted within the sampled groups. All conditions use the same group draw. We use 10,000 replicates with seed 20260918. The two primary model-specific intervals are 97.5% intervals (Bonferroni adjustment for two comparisons); other intervals are descriptive 95% intervals. Both primary estimates exceed ten points and both adjusted intervals exclude zero. Gemma's adjusted lower bound is 9.91, not evidence that its true effect necessarily

Table 8: Secondary candidate-visibility effects within Mixed evidence (visible minus hidden, ordinary year generation). Intervals are descriptive 95% intervals; invalid-answer bounds are not confidence intervals. These contrasts do not replace the primary D.

Model	Mode	Effect [95% CI], points	Invalid-answer bounds
Qwen3.8	Off	-38.12 [-45.25, -30.94]	[-38.12, -37.67]
	Medium	8.97 [3.14, 14.86]	[6.28, 8.97]
Gemma 4	Off	-6.28 [-10.45, -2.21]	[-7.17, -4.93]
	Medium	11.21 [6.79, 15.86]	[6.28, 12.11]

exceeds ten points. These intervals do not account for all historical exploration or variation between repeated generations.

For a linear contrast with positive and negative coefficients, the invalid-answer lower bound assigns unknown successes to zero for positive coefficients and one for negative coefficients; the upper bound reverses those assignments. Valid but incorrect outputs remain incorrect. The primary all-invalid bounds are [42.15,45.29] points for Qwen and [10.76,18.83] for Gemma. They concern unknown outcomes of invalid answers only; changing the budget could also change valid answers. Gemma's lower bound itself has a bootstrap 95% interval of [4.85,16.89]. Thus the point bound above ten does not guarantee a ten-point population effect after accounting for sampling uncertainty.

B.4 Scale and individual-answer changes

For s successes out of n, use $\ell_a(s,n) = \log((s+a)/(n-s+a))$ and compute the same primary contrast after replacing each cell accuracy by this smoothed marginal log-odds. This post-hoc sensitivity uses $a = 0.25, 0.5, 1$ and the same source-group draws. It is not a fitted mixed-effects model. Every primary interval includes zero, so the accuracy-scale finding does not establish a scale-independent interaction. Conversely, these intervals do not establish that a log-odds effect is absent.

The ordinary-generation primary contrast is the sum of the visible-minus-hidden change in Δ_{medium} and the negative change in Δ_{off}. Those terms are 8.97 and 36.32 points on Qwen, versus 11.21 and 5.83 on Gemma. This identity describes observed accuracy components, not causal mediation. In particular, zero medium accuracy on Duplicate prevents interpreting a large interaction as successful resolution of repeated misinformation.

B.5 Later GLM extension: coverage, failures, and scoring

GLM-4.6V-Flash adds a smaller, 9B-class model family, not a parameter-matched control. It uses the same 223 temporal tasks and 256 chart images as the preceding models, with the eight ordinary-generation cells in each task. It does not extend the syntax-constrained, multiple-choice, repeated-correct-chart, CMC, or pet comparisons. The original two-model primary tests are unchanged; this later extension and its intervals are descriptive, not a third member retroactively added to that testing family.

Runtime. The snapshot revision is `411bb4d77144a3f03accbf4b780f5acb8b7cde4e`. GLM runs unmodified BF16 weights on one GPU with vLLM 0.27.1, a 16,384-token context, automatic KV-cache dtype, enabled prefix caching and chunked prefill, and the native image processor. Requests are serial, with temperature zero, seed 20260921, `include_reasoning=true`, and `enable_thinking=true/false` in the chat template. The reasoning parser is `glm45`; no medium-effort parameter is set. The 8,192-token output cap includes reasoning and final answers. These are the study's deterministic settings, not the publisher's sampled leaderboard recipe (Z.ai, 2025); native modes and runtime details differ across models.

Temporal chronology and budget sensitivity. The initial GLM engineering run failed its answer-validity threshold. Later collection retained all requests and scored final-answer success under a fixed limit; it did not retroactively pass that threshold. Each complete budget arm has 1,784 requests, with the same prompts, final-year parser, and source-group bootstrap as the ordinary temporal core. The 8,192- and 9,500-token collections have identical final strings, parsed answers, termination labels,

Table 9: All controlled-run output failures and mean completion tokens (including reasoning). Each row has 223 tasks. Other denotes invalid but not token-capped. Empty is reported separately and can overlap capped; these columns must not be added.

Model	Interface	Mode	Evidence	Cap	Other	Empty	Tokens
Qwen3.8	Hidden free	off	Duplicate	0	0	0	5.0
			Mixed	0	0	0	5.0
		medium	Duplicate	0	0	0	97.4
			Mixed	6	0	6	1209.4
	Visible free	off	Duplicate	0	0	0	5.1
			Mixed	0	1	0	5.1
		medium	Duplicate	0	0	0	104.2
			Mixed	0	0	0	315.7
	Hidden syntax	off	Duplicate	0	0	0	5.0
			Mixed	0	0	0	5.0
		medium	Duplicate	0	0	0	96.4
			Mixed	6	0	6	1208.5
	Visible syntax	off	Duplicate	0	0	0	5.0
			Mixed	0	0	0	5.0
		medium	Duplicate	0	0	0	103.2
			Mixed	0	0	0	314.7
	MCQ	off	Duplicate	0	0	0	2.0
			Mixed	0	0	0	2.0
		medium	Duplicate	0	0	0	101.4
			Mixed	0	0	0	346.1
Gemma 4	Hidden free	off	Duplicate	0	0	0	5.0
			Mixed	0	2	0	6.4
		medium	Duplicate	0	0	0	156.4
			Mixed	11	0	11	1653.5
	Visible free	off	Duplicate	0	0	0	5.0
			Mixed	0	3	0	9.3
		medium	Duplicate	0	0	0	218.4
			Mixed	2	0	2	917.0
	Hidden syntax	off	Duplicate	0	0	0	5.0
			Mixed	0	0	0	5.0
		medium	Duplicate	0	0	0	156.4
			Mixed	11	0	11	1653.5
	Visible syntax	off	Duplicate	0	0	0	5.0
			Mixed	0	0	0	5.0
		medium	Duplicate	0	0	0	218.4
			Mixed	2	0	2	917.0
	MCQ	off	Duplicate	0	0	0	2.0
			Mixed	0	0	0	2.0
		medium	Duplicate	0	0	0	214.3
			Mixed	0	0	0	925.9

Table 10: Post-hoc primary visibility contrasts on smoothed marginal log-odds, with paired 95% intervals. Every interval includes zero.

Model	Pseudocount a	Contrast [95% CI]
Qwen3.8	0.25	1.23 [-1.27, 2.73]
	0.5	1.29 [-0.62, 2.70]
	1.0	1.39 [-0.01, 2.65]
Gemma 4	0.25	0.37 [-1.52, 1.23]
	0.5	0.45 [-0.92, 1.23]
	1.0	0.54 [-0.39, 1.21]

Table 11: Controlled-run off-to-medium transitions. W and C denote wrong and correct; invalid answers count as wrong.

Model	Candidate / response	Evidence	W–C	C–W	W–W	C–C
Qwen3.8	Hidden / generate	Duplicate	0	6	217	0
		Mixed	54	16	43	110
	Visible / generate	Duplicate	0	2	221	0
		Mixed	143	0	39	41
	Hidden / syntax	Duplicate	0	6	217	0
		Mixed	54	16	43	110
	Visible / syntax	Duplicate	0	2	221	0
		Mixed	137	0	39	47
	Visible / MCQ	Duplicate	0	0	222	1
		Mixed	173	2	34	14
Gemma 4	Hidden / generate	Duplicate	0	2	221	0
		Mixed	59	9	48	107
	Visible / generate	Duplicate	0	1	222	0
		Mixed	89	0	32	102
	Hidden / syntax	Duplicate	0	2	221	0
		Mixed	59	9	48	107
	Visible / syntax	Duplicate	0	1	222	0
		Mixed	89	0	32	102
	Visible / MCQ	Duplicate	0	2	221	0
		Mixed	184	0	37	2

and capped-answer membership on all 1,784 pairs. They are budget checks, not repeated-seed trials; equality of final outputs does not imply identical reasoning streams. Both retain 25 capped, empty final answers. The 8,192-token contrast is -5.83 points, with descriptive 95% interval $[-13.72, 2.26]$ and 97.5% interval $[-14.78, 3.57]$. No sampling interval establishes a negative population effect.

Chart chronology and wrapper rule. The chart prompt and eight wrong-evidence conditions follow Appendix C. An initial strict-number collection stopped after 96 formal requests when a format check found GLM box wrappers; it is excluded, not spliced into the complete run. Before restarting, a gold-blind rule was fixed to accept either the original complete numeric answer or one complete wrapper: an optional literal `The answer is` prefix, then `<|begin_of_box|>`, one answer, and `<|end_of_box|>`. The inner string must pass the original numeric parser; the whole response must match. Extra explanations, multiple boxes, and arbitrary number extraction are not accepted. This model-specific rule is not parser-identical to the Qwen/Gemma comparison. The amended run completed 64 engineering requests plus eight repeated probes, then 2,048 formal requests plus eight repeats. Structural and recorded repeat checks pass; repeats do not enter accuracy denominators. All six invalid formal outputs are retained, including one capped output: the cap is not a seventh failure.

Table 12: GLM full-denominator outcomes. Capped answers are a subset of invalid answers, not an additional category. Tokens include reasoning and final output. The temporal 9,500-token run has identical correctness and failure counts.

Task	Mode	Evidence	Candidates	Correct	Invalid	Cap	Tokens
Temporal	Off	Duplicate	Hidden	2/223	0	0	5.3
			Visible	2/223	0	0	5.3
		Mixed	Hidden	7/223	0	0	5.3
			Visible	7/223	0	0	5.3
	On	Duplicate	Hidden	0/223	0	0	62.3
			Visible	1/223	0	0	62.6
		Mixed	Hidden	96/223	20	20	1390.1
			Visible	84/223	5	5	539.6
ChartQA	Off	Duplicate	Hidden	221/256	1	0	8.2
			Visible	221/256	1	0	7.1
		Mixed	Hidden	209/256	1	0	7.8
			Visible	219/256	1	0	7.2
	On	Duplicate	Hidden	240/256	1	1	331.6
			Visible	249/256	0	0	174.9
		Mixed	Hidden	237/256	1	0	373.3
			Visible	249/256	0	0	202.2

The chart contrast is -2.73 points, with descriptive 95% interval $[-7.03, 1.17]$ and 97.5% interval $[-7.42, 1.95]$.

Separate publication sensitivity. Applying the signed all-invalid assignments to the saved GLM rows gives logical D bounds $[-14.80, -3.59]$ for temporal tasks and $[-3.91, -1.56]$ for charts. These additional bounds were computed during manuscript preparation, without replacing the frozen scores. They describe the observed sample under arbitrary correctness of invalid answers, not sampling uncertainty or a prespecified confirmatory endpoint. Neither a negative point estimate nor two intervals containing zero demonstrates a temporal–chart difference, equivalence, or a model-independent mechanism.

B.6 Exploratory controls on candidate content and statement values

Question and scope. Does the weaker Qwen off baseline depend on which years appear in the candidate list, rather than visibility alone? Two complementary interventions examine reference-year coverage and overlap with incorrect statement years. Both use the same 64 previously examined temporal source groups, selected by a fixed hash rather than model outcomes from 222 eligible tasks; eight disjoint groups serve only engineering checks. One of the original 223 tasks is excluded because its date-valued answers collapse to the same year. These are exploratory follow-ups, not independent confirmation. All predictions are newly collected; the second intervention does not pool old outputs.

Protocol shared by the two interventions. Each task uses a four-digit reference year g, two distinct incorrect years, and a filler year. Duplicate repeats one incorrect statement; Mixed combines two statements with different incorrect years. Visible lists contain four distinct years. The prompt adds: "Any candidate hints provided may be incomplete. You may answer with a year not listed among the hints." The required output is still one four-digit year, with no list-constrained decoding. The same unmodified BF16 Qwen and Gemma services use temperature zero, seed 20260826, and native reasoning off/on, with output caps 6,144 and 8,192 respectively. Caps are fixed within, not matched across, models. Reference answers guide experimental construction, not a model-side selection rule. This wording and year normalization differ from the core protocol; results are not pooled with that comparison or interpreted as the isolated effect of a prompt sentence.

Reference coverage and wrong-year overlap. The first intervention crosses reference presence ($G = 1/0$) with presence of the original incorrect-year pair ($O = 1/0$), plus a hidden-list condition. Replacements use one fixed year-generation scheme and seed. Crossing these five arms with evidence condition and reasoning mode gives 20 requests per task (1,280 per model). Table 13 reports every cell. With the reference visible under Mixed, removing the incorrect years raises Qwen off accuracy by 23.44 points and on accuracy by 4.69 points. The difference in reasoning gains is 18.75 points, with paired 95% interval $[4.69, 34.38]$; Gemma's corresponding estimate is 1.56 $[-3.13, 6.25]$. Qwen on also answers 87.50% correctly when the reference is absent but incorrect years remain. Thus recognizing a supplied correct option is not necessary for all correct answers here; neither this observation nor the interventions establish reliance on pixels rather than prior knowledge.

Table 13: Candidate-content exploration, 64 previously examined source groups. Accuracies are percentages, with denominator 64 in every cell. G indicates reference-year presence; O indicates presence of the original wrong-year pair. All visible lists have four years; table-external answers are allowed. *Includes one capped answer, counted as incorrect.

		Qwen3.8		Gemma 4	
Evidence	Condition	Off	On	Off	On
Mixed	Hidden	70.31	87.50	71.88	81.25
	$G1O1$	54.69	85.94	71.88	84.38
	$G1O0$	78.13	90.63	70.31	81.25
	$G0O1$	48.44	87.50	68.75	78.13
	$G0O0$	71.88	82.81	68.75	78.13*
Duplicate	Hidden	14.06	6.25	0.00	0.00
	$G1O1$	15.63	9.38	1.56	0.00
	$G1O0$	31.25	9.38	3.13	3.13*
	$G0O1$	12.50	1.56	1.56	0.00
	$G0O0$	31.25	3.13	1.56	0.00

Crossing statement years with candidate years. Replacing options also changes their numerical identities. To balance their additive effects, the second intervention crosses two statement versions, S_A and S_B, with two candidate lists, C_A and C_B. S_B replaces only the incorrect years in S_A; C_A and C_B contain their respective incorrect-year pairs, plus the same reference and filler in fixed slots. Consequently, the diagonal pairs S_A/C_A and S_B/C_B overlap in incorrect years, whereas the off-diagonal pairs do not. Each statement version and candidate list occurs once on each side of this comparison. All four crosses have equal text-token length within each model. Crossing reasoning mode and evidence condition gives 16 requests per task (1,024 per model).

For a fixed evidence condition, let $a_m(S, C)$ denote percentage accuracy under mode m. Define the overlap effect and its reasoning-associated difference as

$$O_m = \tfrac{1}{2}\big[a_m(S_A, C_A) + a_m(S_B, C_B) - a_m(S_A, C_B) - a_m(S_B, C_A)\big], \qquad J = O_{\text{on}} - O_{\text{off}}.$$

Off recovery after removing overlap is $-O_{\text{off}}$. J **is not the visibility contrast** D**:** it varies candidate content within a fixed evidence condition, rather than comparing the Mixed–Duplicate interaction between hidden and visible interfaces. Qwen/Mixed J is the primary exploratory contrast. Intervals resample the 64 source groups jointly over all cells, with 10,000 draws and seed 20260925; secondary intervals are unadjusted. Averaging the two statement versions does not create 128 independent examples.

What changes, and what remains unresolved. Balanced over statement versions, Qwen/Mixed off accuracy is 53.13% with overlap and 81.25% without; on accuracy is 83.59% and 89.06%. Hence overlap yields a larger reasoning gain (30.47 versus 7.81 points) despite lower final accuracy. The gain difference $J = 22.66$ has interval $[12.50, 33.59]$; the off recovery is 28.13 $[18.75, 37.50]$. Both statement orientations have positive point estimates (Table 15). Gemma's balanced $J = 3.13$ has interval $[-1.56, 8.59]$, and off recovery is 0.78 $[-3.13, 4.69]$. These intervals do not establish equivalence; a significant Qwen effect and a non-significant Gemma effect do not themselves test a model difference. Duplicate on accuracies remain low, so removing overlap is not a general repair.

Table 14: Crossed statement and candidate years on the same 64 source groups. Accuracies are percentages; every cell has denominator 64. S_A/C_A and S_B/C_B share the incorrect statement years; the other two do not. The correct year and filler remain visible in fixed slots. *Includes one capped answer, counted as incorrect.

		Qwen3.8		Gemma 4	
Evidence	Condition	Off	On	Off	On
Mixed	S_A/C_A	54.69	85.94	71.88	84.38
	S_A/C_B	78.13	90.63	70.31	81.25
	S_B/C_A	84.38	87.50	68.75	75.00
	S_B/C_B	51.56	81.25	65.63	76.56
Duplicate	S_A/C_A	15.63	9.38	1.56	0.00
	S_A/C_B	31.25	9.38	3.13	3.13*
	S_B/C_A	29.69	0.00	3.13	1.56
	S_B/C_B	17.19	4.69	0.00	0.00

Table 15: Crossed-content contrasts in percentage points, with paired source-group 95% bootstrap intervals (10,000 draws). Only Qwen/Mixed J is the primary contrast of this exploration; other intervals are unadjusted secondary descriptions. None establishes independent confirmation or a direct model-difference test.

Contrast	Qwen3.8	Gemma 4
Mixed: balanced J	22.66 [12.50, 33.59]	3.13 [-1.56, 8.59]
Mixed: off recovery	28.13 [18.75, 37.50]	0.78 [-3.13, 4.69]
S_A: gain difference	18.75 [4.69, 34.38]	1.56 [-3.13, 6.25]
S_B: gain difference	26.56 [14.06, 40.63]	4.69 [-3.13, 12.50]
S_A: off recovery	23.44 [14.06, 34.38]	-1.56 [-4.69, 0.00]
S_B: off recovery	32.81 [21.88, 43.75]	3.13 [-4.69, 10.94]
Duplicate: balanced J	16.41 [9.38, 24.22]	0.00 [-3.13, 3.13]
Duplicate: off recovery	14.06 [7.81, 21.09]	2.34 [0.00, 5.47]

The cross balances additive statement- and candidate-identity effects, not every nonlinear interaction between years, knowledge, and context. One replacement scheme, one template, and a reused cohort limit generalization. No internal information flow is measured. Using known incorrect values to construct controls does not supply a deployable method for deciding which values to remove when the correct source is unknown.

Failures and integrity checks. Qwen has no invalid exploration outputs. Gemma has two capped outputs in the first intervention and one in the crossed intervention, all retained as incorrect; none affects either Mixed primary comparison. The crossed cap is in S_A/C_B, Duplicate/on. A strict full-string year check leaves all cell counts unchanged. The invalid-content bounds for Gemma's Duplicate J are $[-0.78, 0.00]$ points; these sample-level logical bounds differ from its bootstrap interval. Within-run repeats pass. Across collections, the 576 requests per model that retain the original statement version have identical input hashes, parsed answers, stop labels, and cap status. This consistency check is not independent replication.

C Prospective chart and label-vocabulary follow-ups

C.1 ChartQA: selection and controlled prompts

Unless specified otherwise, the collection and scoring details below concern the original Qwen/Gemma comparison. GLM uses the same chart cohort and wrong-evidence core; its later collection, wrapper amendment, and complete cells are in Appendix B.5. The chart cohort was selected before observing its model outputs, with one question per image and 128 questions from each of the human-authored and machine-generated test subsets. Eligible references are single positive numbers. Fixed rules exclude year/time questions and ratio, fraction, probability, and related ques-

tions with potential unit ambiguity. Explicit percentage questions require all candidates to lie in $(0, 100]$. Distractors come from positive validation answers, match decimal precision and order of magnitude, and have non-overlapping 5% acceptance intervals. The visible list contains the reference and three distractors; reference position rotates deterministically. This is an intervention supplying answer information, not independently retrieved evidence.

Decoded RGB hashes and 64-bit difference hashes (Hamming distance at most four) exclude images overlapping cached MMMC, MMKC, CMC, M4FC, the eight engineering charts, or the entire cached ChartQA validation split. The same checks exclude overlap within the new cohort. Eligible examples are selected by a fixed salted hash, without correctness screening. These checks do not establish semantic-source independence, absence from pretraining, or human-verified label quality. The experiment's instruction explicitly prioritizes chart evidence; it does not identify a trustworthy source when source reliability is unknown.

Each image has 12 conditions: two candidate settings, two reasoning modes, and three statement conditions. In wrong Duplicate, both statements use the first distractor; wrong Mixed uses two distinct distractors; the aligned reference repeats the correct value. The latter is reported separately and does not enter the primary D. The exact text template, following the image in the same user message, is:

> Answer the question using the chart. Two accompanying text statements may be incorrect.
> Question: [question]
> Text statement 1: The answer is [value 1].
> Text statement 2: The answer is [value 2].
> *[Visible only:]* Candidate answers: [four comma-separated values].
> Return only a single number as your final answer, without units or explanation. Use the units requested by the question or shown in the chart; do not convert percentages to fractions or expand the displayed unit scale.

Within a model, the candidate block is the only hidden/visible prompt difference. Models use unmodified BF16 weights, serial requests, temperature zero, seed 20260826, and the same native reasoning settings as the controlled temporal study. Qwen has a 6,144-token output cap and Gemma an 8,192-token cap; this is not a common-limit comparison. The valid Qwen collection used the recorded GPU6 service migration, not the earlier interrupted service. Each model completed 96 engineering requests plus 12 repeated probes, followed by 3,072 formal requests plus 12 repeats. Repeats check parsed outcomes and termination status, not identity of every reasoning token. Interrupted runs and engineering examples are not pooled into the scientific denominator.

C.2 Chart scoring, uncertainty, and negative result

The original parser accepts one complete numeric answer, not numbers extracted from prose. Relative error at most 5% is correct; exact numeric equality is a sensitivity. Capped, empty, and unparsed answers count as failures. Whole images are resampled jointly across all conditions in 10,000 paired bootstrap replicates with seed 20260918. Each model's original positive replication criterion requires $D \geq 10$ points and a positive lower endpoint of the two-sided 97.5% interval. Neither model meets it (Table 17). The two primary intervals are adjusted within this follow-up's pair of tests, not across every historical or supplementary comparison.

The all-invalid logical bounds assign correctness to every invalid answer according to its sign in D; they are not confidence intervals. Qwen has one capped and one other invalid output; Gemma has 20 capped and 88 other invalid outputs. The latter bounds still permit a large positive answer-content contrast, even though the strict-score estimate is small. Correct-reference cells are retained in Table 16; low strict accuracy in Gemma's hidden/off reference condition also reflects numeric formatting failures, not necessarily obedience to incorrect evidence.

A separate post-hoc rule accepts complete, legally grouped thousands and a complete number with a percent suffix only when the question explicitly requests percent, percentage, or %. It changes no scale or units, extracts no number from prose, and never uses the reference to decide parse validity. It admits 18 Gemma percentage outputs and one grouped number, all correct under the original tolerance. Four occur in hidden/off wrong Duplicate, four in hidden/off wrong Mixed, and 11 in the aligned reference. The primary D is unchanged by cancellation, not because every cell is unchanged. Gemma retains 89 invalid answers, including the 20 caps; Qwen has no newly accepted answers.

Table 16: All prospective ChartQA cells. Correct counts use the original 5% relative tolerance and strict numeric parser; denominator 256 in every row. Invalid includes capped outputs, not a disjoint category.

Model	Candidates	Mode	Statements	Correct	Acc. (%)	Invalid	Cap
Qwen3.8	Hidden	off	Wrong duplicate	236	92.19	0	0
			Wrong mixed	235	91.80	1	1
			Correct duplicate	245	95.70	0	0
		medium	Wrong duplicate	246	96.09	0	0
			Wrong mixed	246	96.09	0	0
			Correct duplicate	248	96.88	0	0
	Visible	off	Wrong duplicate	244	95.31	1	0
			Wrong mixed	241	94.14	0	0
			Correct duplicate	250	97.66	0	0
		medium	Wrong duplicate	251	98.05	0	0
			Wrong mixed	252	98.44	0	0
			Correct duplicate	252	98.44	0	0
Gemma 4	Hidden	off	Wrong duplicate	217	84.77	24	0
			Wrong mixed	218	85.16	24	0
			Correct duplicate	210	82.03	33	0
		medium	Wrong duplicate	240	93.75	6	4
			Wrong mixed	240	93.75	7	7
			Correct duplicate	249	97.27	0	0
	Visible	off	Wrong duplicate	244	95.31	1	0
			Wrong mixed	242	94.53	3	0
			Correct duplicate	244	95.31	0	0
		medium	Wrong duplicate	249	97.27	3	3
			Wrong mixed	248	96.88	6	5
			Correct duplicate	253	98.83	1	1

Table 17: ChartQA candidate-visibility contrasts D. Original intervals are the 97.5% primary intervals; numeric-format intervals are post-hoc descriptive 97.5% intervals. Invalid-answer logical bounds are not confidence intervals.

Model	Parser	D [97.5% CI], points	Invalid bounds
Qwen3.8	Original	1.17 [-1.56, 3.91]	[1.17, 1.95]
	Numeric format	1.17 [-1.56, 3.91]	[1.17, 1.95]
Gemma 4	Original	0.78 [-3.52, 5.08]	[-13.67, 15.23]
	Numeric format	0.78 [-3.52, 4.69]	[-12.11, 13.67]

This rule does not replace the frozen primary decision or resolve all semantic alternatives. Exact-numeric and marginal log-odds sensitivities, tokens, and paired answer transitions remain in the linked machine-readable results.

The chart result restricts transfer of the large temporal interaction under the tested protocol. High off accuracy, different prompts and distractors, and image-supported answers vary together. It does not establish which difference explains non-replication, equivalence to zero, or a general visual-evidence mechanism.

C.3 Pet recognition: new images and trainval vocabulary

The fixed vocabulary contains all 37 Oxford-IIIT Pet breed names, constructed solely from the official trainval metadata by removing the image number, replacing underscores with spaces, and lowercasing. A fixed salted-hash ordering selects five eligible test images per breed (185 in total). Selection is independent of model outputs. The eight engineering images come from trainval, with four cat

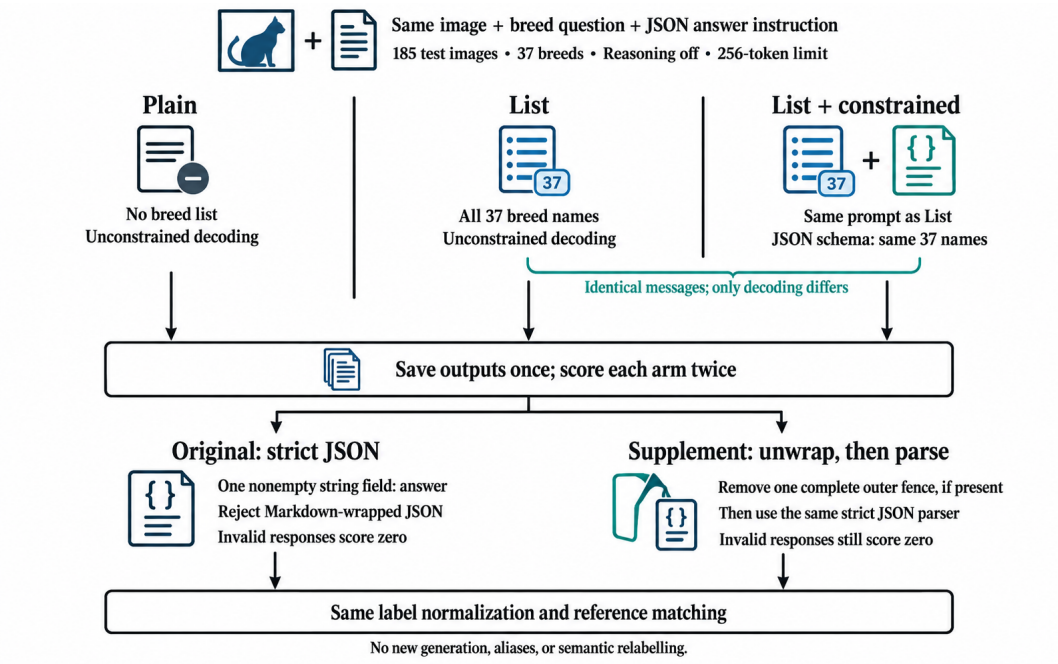


Figure 8: Label exposure, decoding, and scoring in the pet-recognition control. The list contains all 37 trainval breed names, including the reference. List and List+constrained receive identical messages; only decoding differs. Both scoring rules use the same saved outputs and label normalization, without new generation or semantic relabelling. The supplement permits one complete outer Markdown fence; it was chosen after Qwen's results and fixed before Gemma's collection, and does not replace either original primary score. All arms disable reasoning; this is not a conflict or reasoning-mode comparison. Icons are schematic, not benchmark images or model outputs.

and four dog breeds, and are excluded from formal sampling. Image-hash and near-duplicate checks include the prospective chart study history index, all cached ChartQA validation/test images, and within-cohort comparisons. Class coverage is closed-set and complete; the experiment does not test new breeds, missing-reference lists, or pretraining-unseen animals.

The plain prompt, following the image, is:

> What is the breed of the main cat or dog in this image?
> Return only a JSON object with one string field "answer" containing the breed name, without explanation.

List and guided prompts insert `Task breed vocabulary:` followed by the same alphabetical list, one hyphen-prefixed name per line, before the final instruction. Their messages are identical. Guided decoding alone adds a JSON schema: an object with exactly one required string property `answer`, whose enumeration is the same list, with additional properties disallowed. The prompt gives no dataset name, filename, per-image class annotation, correct-class indicator, or image-specific breed hint. The full vocabulary nevertheless includes the reference breed along with all other classes. The image itself is sent without segmentation or bounding-box annotations.

Both models disable reasoning and use temperature zero, seed 20260826, a 256-token cap, and serial requests. Arms are interleaved by a fixed hash. Each model completes 24 engineering requests plus 24 repeats and 555 formal requests plus nine repeats. Collection, request, reasoning-mode, zero-retry, and schema checks pass. No formal output reaches the cap. Each arm/image has one deterministic generation, not repeated-seed estimates.

Table 18: Pet recognition under both parsers (185 images per arm). All arms disable reasoning and use a 256-token cap. Wrapper removal changes only acceptance of one complete outer fence, not the generated answers or normalization.

Model	Parser	Arm	Correct	Acc. (%)	Invalid	Cap
Qwen3.8	Strict JSON	plain	81	43.78	59	0
		list	45	24.32	132	0
		guided	157	84.86	0	0
	Outer fence removed	plain	104	56.22	21	0
		list	158	85.41	0	0
		guided	157	84.86	0	0
Gemma 4	Strict JSON	plain	0	0.00	185	0
		list	0	0.00	185	0
		guided	164	88.65	0	0
	Outer fence removed	plain	131	70.81	0	0
		list	164	88.65	0	0
		guided	164	88.65	0	0

Table 19: Pet paired comparisons. Intervals are 97.5%; supplementary wrapper-adjusted intervals are descriptive. Repairs/regressions count wrong-to-correct and correct-to-wrong transitions. Zero observed difference is not population equivalence.

Model	Parser	Comparison	Change [97.5% CI], points	Repair / regress
Qwen3.8	Strict JSON	List − plain	-19.46 [-27.03, -11.35]	20 / 56
		Guided − list	60.54 [54.59, 67.03]	112 / 0
	Outer fence removed	List − plain	29.19 [22.16, 35.68]	66 / 12
		Guided − list	-0.54 [-1.62, 0.00]	0 / 1
Gemma 4	Strict JSON	List − plain	0.00 [0.00, 0.00]	0 / 0
		Guided − list	88.65 [84.32, 92.43]	164 / 0
	Outer fence removed	List − plain	17.84 [12.43, 23.24]	38 / 5
		Guided − list	0.00 [0.00, 0.00]	0 / 0

C.4 Two scoring rules, with their chronology retained

The original parser requires a JSON object with exactly one nonempty string `answer`; duplicate keys and extra fields are invalid. Predictions and references both receive NFKC, whitespace collapse, and case folding, without punctuation or alias repair. Invalid responses remain in the full 185-image denominator. The original main test is list minus plain: point estimate at least five points and positive lower endpoint of the 97.5% interval. Both models fail that test under strict parsing.

After observing Qwen's outputs, we fixed a supplementary parser that additionally accepts one complete outer triple-backtick fence, either unlabelled or labelled `json`. The entire response, except surrounding whitespace, must be that fence; prose outside, multiple or nested blocks, other language labels, malformed JSON, and empty objects remain invalid. The inner JSON parser and normalization are unchanged. The supplement and implementation hashes were sealed before Gemma's first output in this study. It is post-hoc for Qwen and prospectively specified as a supplementary endpoint for Gemma, but it was informed by the other model's outcomes and is not the original primary endpoint for either model. We preserve both scores rather than retroactively change the primary test.

Qwen has 38 fenced plain and 132 fenced list outputs; 21 plain empty objects remain invalid after wrapper removal. Gemma has 185 fenced outputs in each unconstrained arm, all parseable after the fixed removal rule. The strict-score difference from guided decoding therefore cannot be attributed wholly to recognizing the image. Under wrapper-adjusted scoring, the paired list benefit repairs/regresses 66/12 Qwen answers and 38/5 Gemma answers. Qwen's remaining-invalid logical range for list minus plain is [17.84,29.19] points. Gemma has no remaining invalid answers. Guided

decoding yields one fewer correct Qwen answer and the same number of correct Gemma answers as the list arm; equal correctness does not mean every output string is identical or that population accuracy is equivalent.

Intervals use 10,000 paired bootstrap replicates, seed 20260920, sampling five images with replacement within each of the 37 fixed classes and sharing each draw across arms. They target the uniform average over these classes, not new-category generalization or run-to-run variation. Wrapper-adjusted 97.5% intervals are descriptive, not additional confirmatory multiplicity control. Gemma's degenerate zero-difference interval cannot establish equivalence: its normalized list/guided strings differ on two images despite identical correctness states. Qwen's strings differ on one image. Raw-string matching remains in the retained results, not the primary comparison.

Prior exploration and scope. The design was motivated by an earlier label-list study on 96 already examined development objects in 48 pairs. That exploratory study is not counted as an independent confirmation here. The pet follow-up uses new images and a taxonomy fixed from trainval; it still does not isolate naming from recognition or independently adjudicate semantic alternatives.

D Cases from saved model outputs

These post-hoc illustrations are not independent tests or estimates of typical behavior. For each chosen correctness pattern, the lexicographically first matching task is shown. Selection is outcome-dependent: the patterns were chosen to explain the aggregate results. Counts below refer to the complete cohorts. No model is rerun; only final answers, not reasoning traces, are displayed. References are benchmark annotations.

D.1 Two different origins of an enlarged interaction

Both tasks ask: "When was the brand associated with the logo established?" The question image remains in every condition. Duplicate repeats the first incorrect statement; Mixed replaces its second copy with the second incorrect statement.

Qwen3.8. `logo_knowledge:14:time`; reference 1997. Incorrect statements: (1) "Netflix was created in 2000. The creator of Netflix is Reed Hastings. Netflix is the brand of streaming service." (2) "Netflix was created in 1995. The creator of Netflix is Reed Hastings. Netflix is the brand of streaming service." Visible candidates: A: 1997; B: 1998; C: 1995; D: 2000. All eight outputs are valid for 74/223 tasks with this exact correctness pattern, including this example.

Gemma 4. `logo_knowledge:10:time`; reference 1926. Incorrect statements: (1) "Mercedes-Benz was created in 1950. The creator of Mercedes-Benz is Karl Benz and Gottlieb Daimler. Mercedes-Benz is the brand of luxury vehicles." (2) "Mercedes-Benz was created in 1900. The creator of Mercedes-Benz is Karl Benz and Gottlieb Daimler. Mercedes-Benz is the brand of luxury vehicles." Visible candidates: A: 1900; B: 1926; C: 1950; D: 1935. All eight outputs are valid for 14/223 tasks with this exact correctness pattern, including this example.

Table 20: Saved final answers, with only outer whitespace omitted. Bold indicates reference matching. The examples are different tasks, not a controlled between-model comparison.

		Qwen3.8: reference 1997		Gemma 4: reference 1926	
Candidates	Mode	Duplicate	Mixed	Duplicate	Mixed
Hidden	Off	2000	**1997**	1950	1900
	Medium	2000	**1997**	1950	1900
Visible	Off	2000	D. 2000	1950	1900
	Medium	2000	**1997**	1950	**1926**

For Qwen, candidates change Mixed/off from correct to incorrect while leaving all medium correctness states unchanged. Its per-task interaction grows through the off baseline alone. For Gemma, the off answers remain incorrect and only the visible/Mixed medium answer becomes correct. The same sign of interaction change can thus have different cell-level origins. This is an arithmetic illustration, not an internal mechanism. The incorrect statements above are benchmark constructions, not assertions of real-world truth.

D.2 A label-list repair and a regression under two scoring rules

Both Gemma cases use the question and 37-label list in Appendix C. Their list outputs fail strict JSON parsing but become scorable after removing one complete outer fence. The saved text below retains its line breaks and Markdown fences.



miniature_pinscher_82



staffordshire_bull_terrier_36

Figure 9: Full-frame Oxford-IIIT Pet images for the repair (left) and regression (right) (Parkhi et al., 2012). Distributed under CC BY-SA 4.0; image copyrights remain with the original owners. Source: `https://www.robots.ox.ac.uk/~vgg/data/pets/`. Selected illustrations, not new evaluations.

List repair: 38/185 matching tasks. `pets_full_005`. Reference: miniature pinscher.
Plain: strict incorrect; adjusted incorrect.

```
```json
{"answer": "Pharaoh Hound"}
```
```

List: strict incorrect; adjusted correct.

```
```json
{"answer": "miniature pinscher"}
```
```

Guided: strict correct; adjusted correct.

```
{"answer": "miniature pinscher"}
```

List regression: 5/185 matching tasks. `pets_full_022`. Reference: staffordshire bull terrier.
Plain: strict incorrect; adjusted correct.

```
```json
{"answer": "Staffordshire Bull Terrier"}
```
```

List: strict incorrect; adjusted incorrect.

```
```json
{"answer": "american pit bull terrier"}
```
```

Guided: strict incorrect; adjusted incorrect.

```
{"answer": "american pit bull terrier"}
```

The list can help or harm individual answers. Wrapper handling can change whether a response is scorable without changing its breed name. Neither example isolates naming from recognition; aggregate paired results determine the reported benefit.

E Design comparison with the closest option-access studies

The comparison below uses the complete arXiv v2 manuscripts, including appendices, of Raman et al. (2026) and Liu et al. (2026a). Their ICML 2026 and CVPR 2026 publication status was verified separately; this does not imply that the conference final versions or separately hosted supplements were inspected in full.

For Test Exploiters, option timing and controls are in Sections 3–4 and scoring in Appendix C. For ReVeL, Sections 2–3 describe candidate construction and verification; Table 5 and Appendix F show that per-option verification retains option content. Neither inspected version reports the paired Duplicate/Mixed D studied here. This is a bounded comparison of reported designs, not proof of exhaustive priority. Our contrast also does not identify how much gain comes from guessing, elimination, internal evidence arbitration, or particular option cues. Such mechanisms would require further interventions.

Table 21: Specific design differences, not a claim that answer-option effects are new.

Aspect	Test Exploiters	ReVeL	Controlled temporal study here
Task	Text knowledge, mathematics, code, economics	Text/visual QA; evaluation and reinforcement fine-tuning	Image-associated year questions with two incorrect statements
Option access	Options before or after reasoning; options-only and task controls	Open-question rewriting and per-option verification; the latter retains option content	Add/remove the same four-option block, retaining the year instruction
Reasoning	Staged reasoning; Qwen3 also disables thinking in its second stage	Reasoning outputs and training/evaluation formats; no reported matching of our full three-factor design	Native off/medium crossed with hidden/visible candidates and Duplicate/Mixed evidence
Target contrast	Accuracy and effects of option timing/content	Score inflation, verification quality, and training transfer	D: visibility change in the reasoning-by-disagreement interaction
Answer scoring	Task-dependent numeric/symbolic rules on scorable tasks	Rule/LLM hybrid verifier, not raw-string equality	Parsed years; full denominators; invalid and capped outputs fail
Scope	Already establishes option-dependent reasoning performance	Already studies visual option-related inflation and alternative verification	A narrower conflict-conditioned contrast; baseline decomposition and chart non-replication limit interpretation

F Reproduction materials and study provenance

Retained records include task lists, request/configuration hashes, raw outputs, protocol revisions, manifests, and scoring scripts. Supplementary analyses remain separate; interrupted runs are not pooled. Figures use saved values, without new inference.

Table 22: Reader-facing studies and identifiers used only to locate the retained files.

Study	Record identifier	Main retained analysis
Evidence direction	`R13-B`	Absolute accuracy and corrected abstention
Candidate visibility	`R20`	Full cells, paired contrasts, invalid bounds
Chart follow-up	`R21`	Strict scoring and numeric-format sensitivity
Pet vocabulary	`R22`	Strict JSON and separate wrapper supplement
GLM temporal extension	`R23`	End-to-end scores and two output budgets
GLM chart extension	`R23P1`	Fixed wrapper rule and eight-cell results
Candidate content	`R24`	Reference coverage and wrong-year overlap
Crossed year identities	`R24B`	Balanced overlap contrasts on the same cohort

Original temporal, chart, and pet runs retain `full.jsonl`; GLM temporal runs use `rows.jsonl`; candidate-content runs use `pilot.jsonl`. Each has manifests and seals. Publication provenance links values and case selections to source hashes. In GLM chart records, `run_manifest_sha256` denotes the protocol seal, not the collection manifest. The candidate-content collections share 64 source groups without pooling predictions; they are not independent replications.

A local results-level audit candidate exists, not a complete inference-reproduction archive or public release. Portable configurations, output/evaluator packaging, license checks, and end-to-end testing remain. Original weight/image access terms apply. Appendices A–C specify the protocols.