

Asking About Occluded Content Marks Adversarial Multimodal Questions

HyoJung Han*
University of Maryland
hjhan@umd.edu

Atrey Desai*
University of Maryland
adesai10@umd.edu

Ishan Upadhyay
IIT Bombay
ishan.upadhyay.iitb@gmail.com

Tianyi Zheng
UC San Diego
t4zheng@ucsd.edu

Eve Fleisig
Princeton University
efleisig@princeton.edu

Aparna Garimella
Adobe Research, India
garimell@adobe.com

Marine Carpuat
University of Maryland
marine@umd.edu

Jordan Boyd-Graber
Nanyang Technological University
jordan.ying@ntu.edu.sg

Abstract

Adversarial questions, which are harder for machines than for skilled people, show where vision–language models (VLMs) fall short. But finished questions alone cannot show what makes them adversarial: the edits that made them hard are not visible. We study both in multimodal quizbowl: 27 authors wrote questions in an interface logging image revisions and model probes; skilled human teams and AI systems competed on 240 of them. Adapting Item Response Theory to clue-by-clue questions, we estimate how much harder each clue is for machines than for skilled humans. Of 14 question characteristics, only asking about content the image hides is reliably associated with clues being harder for machines (1.17 logits, $q = .003$, seven questions). When authors edit an image between two model probes, 35% of the model’s correct answers turn wrong, even when the needed clues remain; the model then relies on weaker clues or misses connections. Our findings suggest adversarialness depends less on what images show and more on what questions ask respondents to do. More broadly, expert authors probing a model, and their edits, reveal where VLMs trail skilled humans.

1 Introduction

Vision–language models (VLMs) are evaluated on a broad set of tasks, from everyday scenes (Antol et al., 2015; Goyal et al., 2017; Gurari et al., 2018) to expert knowledge (Yue et al., 2024). However, accuracy is agnostic to *which* questions are hard for models but not people: models answer some questions from language priors alone (Agrawal et al., 2018; Chen et al., 2024; Lan et al., 2026) yet fail on visual distinctions trivial to people (Tong et al.,

1 Authoring the author probes a VLM

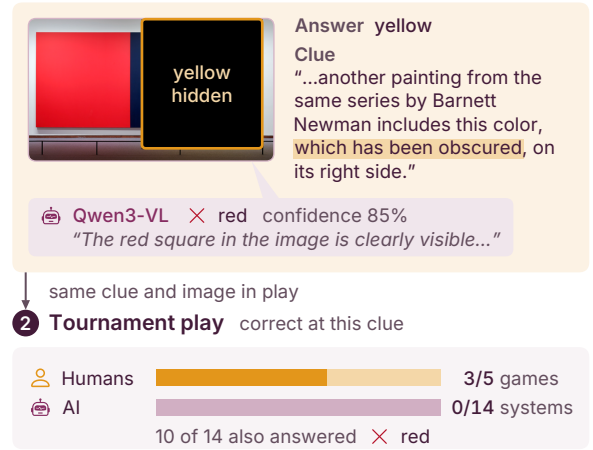


Figure 1: The question asks for the color masked on the right of a Barnett Newman painting (answer: *yellow*). Authors probe a VLM during question creation to test adversarialness; in live competition rounds, most humans (3/5) but no AI systems (0/14) answer correctly.

2024). Comparing skilled humans and machines on the *same* questions locates these gaps along the “jagged frontier” of AI (Dell’Acqua et al., 2026).

Adversarial data collection finds such questions by having humans author and revise examples until a model fails (Nie et al., 2020; Bartolo et al., 2020; Kiela et al., 2021), as is common in VQA (Sheng et al., 2021; Li et al., 2021). However, a model failure alone does not make an example adversarial: the example may be ambiguous, mislabeled, or hard for everyone (Bowman and Dahl, 2021; Northcutt et al., 2021). Following Sung et al. (2025b), we define a question as **adversarial** if it is harder for a machine than for a skilled human, compared with a typical question and after accounting for each group’s overall ability.

Question edit histories reveal what finished ques-

* Equal contribution.

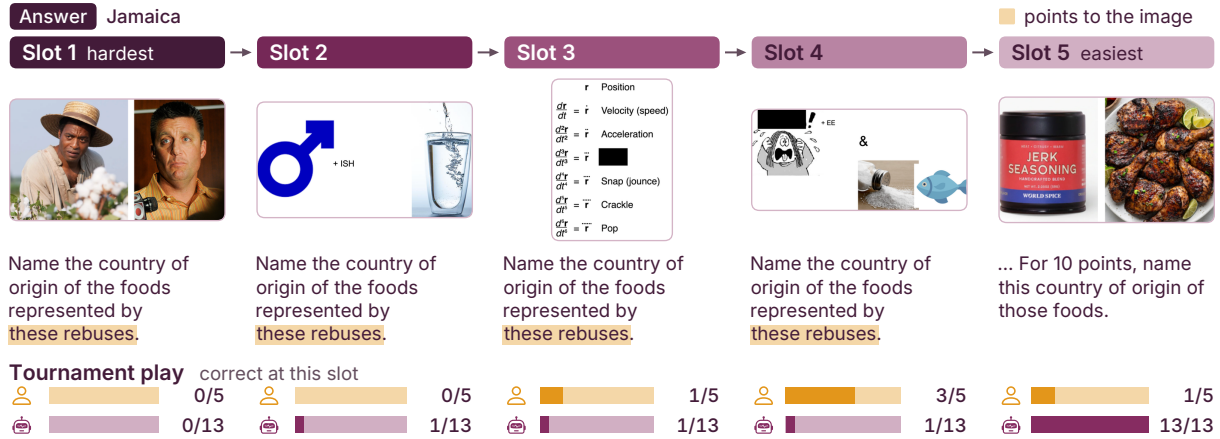


Figure 2: Human teams solve this tossup from its rebuses, while the AI systems need the final, direct clue. Each *slot* pairs an image with a clue (Section 2): the first four images are rebuses for Jamaican foods, with words masked in slots 3 and 4, and the clue repeats until the last slot (clues abridged). Bars count correct answers at each slot, as first human buzzes in five games (person icon) and answers of 13 AI systems (robot icon). Humans answer in all five games, four by slot 4; at most one system answers before slot 5. Because slots grow easier, where each group first answers shows how much evidence it needs.

tions hide: which information authors consider alterable to confuse AIs but not humans. Right intuitions reveal where AIs lag behind humans; wrong ones reveal where authors misjudge AI. Beyond editing text (Jia and Liang, 2017; Wallace et al., 2019), images give authors more freedom to alter the evidence available to a model, such as by masking the color that a question asks for (Figure 1).

We study both the finished questions and their authoring in CHAVIT 2026,¹ a human–AI quizbowl competition (Section 2) whose main questions are revealed in *slots*, each an image with a text clue (Figure 2). Twenty-seven authors wrote multimodal questions using a custom interface (Section 3); skilled human teams and AI systems then competed on 240 of them. We ask:

- RQ1** Which VQA questions are harder for machines than for skilled humans, and what characteristics do they share? (Section 4)
- RQ2** How do authors make VQA questions adversarial, and what do model failures on their probes reveal? (Section 5)

For RQ1, an Item Response Theory (IRT) model that estimates difficulty slot by slot finds that slots asking for content an image implies but does not show, such as masked information, are the clearest marker of where models struggle more than humans. For RQ2, 442 probe–edit–probe pairs from 17 authors show that revisions change

which evidence remains and how it must be used; when answer-relevant clues plausibly survive, the authoring-time model can still fail by prioritizing, connecting, or recognizing them incorrectly.

2 Background: Multimodal Quizbowl

Quizbowl is a trivia competition in which two teams answer questions read aloud by a moderator, interrupting (“buzzing”) as soon as they know the answer (Jennings, 2006). A **tossup** consists of a sequence of *slots*; each slot pairs an image with an (optional) text clue, and all slots share one answer (Figure 2). Each slot’s clue is read incrementally, and its image appears at the point in the reading where it is referenced. A player may buzz at any point, and a correct buzz ends the tossup. Tossups are *pyramidal*: clues grow progressively easier, ending in a “giveaway,” so the position of a correct buzz reveals the minimum information a player needed (Rodriguez et al., 2019). A **bonus** consists of a lead-in sentence followed by three independent parts, each with its own image and answer; unlike tossups, bonuses are neither interruptible nor pyramidal. A game has 20 tossups, and only a team that answers a tossup correctly plays a bonus.

Following Gor et al. (2026), we collect data by hosting quizbowl tournaments (Section 4.1) to attract skilled players who represent the upper end of human ability and need no training on the formats.

¹A fictitious name for anonymity.

3 Authoring Multimodal Questions

In text-based quizbowl, Wallace et al. (2019) display a model’s predictions and their interpretations to authors as they write, then validate the questions in human–computer games. We extend this to images: authors edit images and text, probing a model in the process.

Authoring guidelines. The guidelines list five desiderata: questions should be *multimodal* (not answerable from a single modality), *pyramidal*, *adversarial*, *uniquely identifying* (the clues together point to only one answer), and *well-formed* (free of ambiguity, false presuppositions, and misleading answer types; Boyd-Graber and Börschinger, 2020). These are the norms of *probing* questions, written by someone who knows the answer to test another’s knowledge (Rogers et al., 2023); ambiguity (Min et al., 2020; Stengel-Eskin et al., 2023) and false presuppositions (Kim et al., 2023; Yu et al., 2023) are documented flaws of QA and VQA data. Our guidelines ask that an image be “distinctive enough that an expert familiar with the answer would recognize it, but not so iconic that the image alone gives it away immediately” and encourage transforming images to limit memorization. To this end, we discouraged Wikipedia and public-domain images, which were likely in the models’ training data (Srinivasan et al., 2021; Meyer et al., 2024).

The authoring interface. We built the interface (Figure 3) as a slide-like web editor.² Each question is written in an authoring *session*. Authors create a question from scratch or from a copy of a gallery example, add or reorder slots, upload an image for each slot, and edit it in place: annotation, cropping, rotation, flipping, color adjustments, blur, and filters. Appendix D describes further features.

Model feedback. Three tests expose the model’s responses on the draft; we define a single-slot or buzz test as a *probe*. Replaying the whole tossup after every edit would be slow, and delays of more than a few seconds break the flow of interactive work (Miller, 1968; Card et al., 1991), so the fastest test checks one slot in isolation. The *single-slot test* sends the current slot’s image and the clues so far to a VLM (Qwen3-VL; Bai et al., 2025), which must answer, and returns its answer, confidence,

reasoning, and whether the answer matches the answer line. The *buzz test* plays the whole tossup, letting the model buzz or wait at each slot. The *text-only test* sends the clues without the image, to catch textual shortcuts. For buzz tests and bonuses, authors could select other VLMs, but 94% of logged model runs used Qwen3-VL; these authoring-time probes are separate from the competition’s AI systems. Authors tested 97.1% of the 276 submitted questions at least once (Appendix D).

Logging. The interface logs slot saves, image uploads, tests, model responses, and submissions; authoring is iterative, with a median of 13 slot saves before first submission and 23.9% of submitted questions reopened. Model tests of a slot before and after an image change thus bracket the author’s changes; such probe–edit–probe pairs are RQ2’s unit of analysis (Section 5).

Authors, review, and resulting data. We recruited authors through the CHAVIT call for question writers and provided them with written and video instructions. Authors included experienced quizbowl writers, university students, and researchers; the Ethics Statement describes payment and consent. Five editors reviewed and revised submitted questions and assembled six *packets* of 20 tossups and 20 bonuses each, 240 questions in all (Table 1).

4 Adversarial Questions and Their Characteristics

We first estimate adversarialness below the question level, for each tossup slot and bonus part (Section 4.1), then test its association with 14 question characteristics (Section 4.2): the clearest markers are slots that ask for content the image hides (Section 4.3).

4.1 Measuring Difficulty and Adversarialness

In CHAVIT 2026, teams of up to three skilled players compete alongside two autonomous AI teammates; each team may buzz once per tossup and may consult its AI teammates on bonus parts. Each question is read in up to five of 27 games. Separately, the AI systems submitted to the competition’s shared task are run over the entire question set, abstaining on a tossup until they buzz (Table 1). Overall accuracy (Figure 4) compares groups rather than questions: an accuracy gap may reflect stronger respondents in one group rather

²The interface and analysis code, the questions, and their images are at <https://anonymous.4open.science/r/mmqa-author-release-DBB2/>; URLs for the deployed interface and the call for writers are withheld for anonymous review.

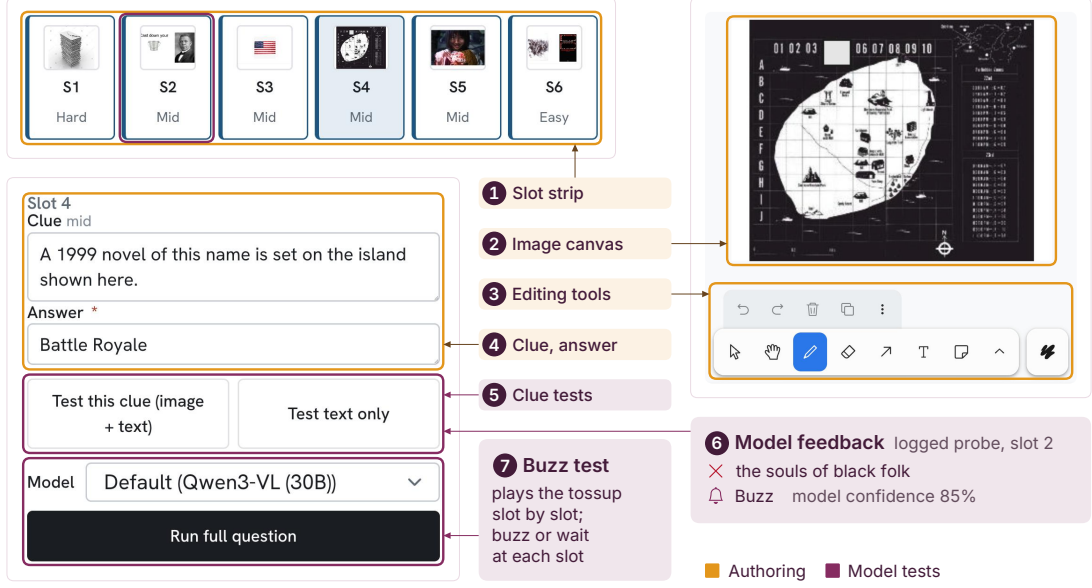


Figure 3: The authoring interface allows authors to view the full question in the slot strip (1), edit visual (2, 3) and text (4) content, and receive model buzz feedback on individual slots (5, 6) and on the full question (7).

than a question-specific machine weakness, and raw accuracy saturates at either difficulty extreme.

Item Response Theory (IRT) models correctness on a latent difficulty scale, accounting for respondents’ overall abilities and a nonlinear relationship between difficulty and accuracy (van der Linden and Hambleton, 2013). The probability that respondent r in group $g_r \in \{H, M\}$ answers item i correctly is

$$P(Y_{ri} = 1) = \sigma(\theta_r - b_i - \mathbb{1}[g_r = M] \delta_i), \quad (1)$$

where σ is the logistic function, θ_r is the ability of respondent r , and b_i is item difficulty for humans (H). As in standard IRT, a respondent is likely to answer correctly when ability θ_r exceeds difficulty b_i . The indicator $\mathbb{1}[g_r = M]$ equals one for machines (M) and zero for humans, so δ_i adds a penalty ($\delta_i > 0$) or boost ($\delta_i < 0$) for machines compared with the human pool. We define the **human-machine difficulty difference** δ_i as a measure of *differential item functioning* (DIF; Holland and Wainer, 1993) and use it as our human-grounded measure of **adversarialness**. Because the data cannot separate group differences in ability from differences in item difficulty, we center the difficulty differences, $\delta_i^c = \delta_i - \bar{\delta}$, so that positive values indicate above-average machine-relative difficulty.

Slot-wise IRT. A tossup unfolds slot by slot, so we propose **slot-wise IRT** to estimate difficulty and adversarialness at each slot. A human *episode*

is one tossup in one game, and its event is the first buzz by a human on either team. All human episodes share one ability and buzz threshold, representing the pool of skilled players: with at most five games per tossup, team ability cannot be separated from slot difficulty. With item parameters b_{is} and δ_{is} for slot s of tossup i , the probability that respondent r first buzzes during slot s , given no earlier buzz and an active opportunity, is

$$P(B_{ris}=1 \mid B_{ri,<s}=0) = \sigma(\theta_r - b_{is}^{g_r} - c_r), \quad (2)$$

where $b_{is}^H = b_{is}$ and $b_{is}^M = b_{is} + \delta_{is}$, c_r is a buzz threshold, and a buzz is correct with probability $\sigma(\theta_r - b_{is}^{g_r})$. Difficulty thus shapes both when a buzz occurs and whether it is correct, whereas the threshold affects only timing: a respondent with a high threshold waits for more clues (for a system, a higher confidence cutoff; Appendix A.1).

The two groups are observed differently. For humans, we see buzzes only until someone answers: in 153 of 490 episodes, an AI teammate answers correctly before any human buzzes, so the episode shows no human buzz during the slots already read but gives no human answer (it is *censored*). For machines, we see every slot: the shared-task systems saved an answer at each one, so the *dense* outcome uses their correctness at every slot, whereas the *strict* outcome uses only their first buzzes, as for humans (Appendices A.1–A.5). Dense is primary because a system’s first buzz reflects its confidence cutoff as much as the slot and because it uses 7,558

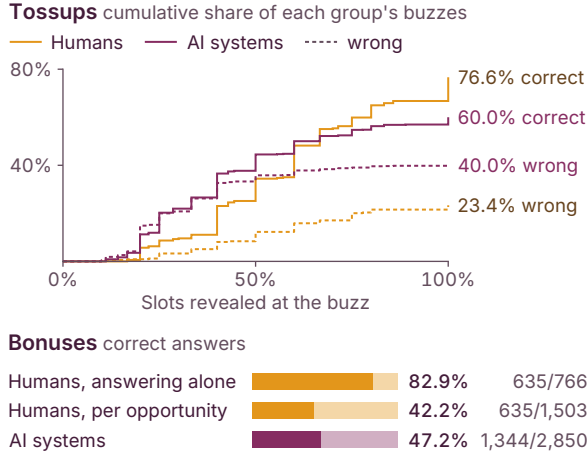


Figure 4: Humans buzz later than the AI systems (median 60% of slots revealed versus 33%) but are more often correct (76.6% versus 60.0%). **Tossups**: cumulative share of each group’s buzzes that are correct (solid) or wrong (dashed), by the share of slots revealed at the buzz; human data are the first human buzz in each game (334 buzzes), and AI data are each of the 14 systems’ first buzz per tossup (1,485). With every slot revealed, the systems answer 95.2% correctly, and the best system is correct on 83.5% of its buzzes. **Bonuses**: a team could consult its AI teammates (612 opportunities) or abstain (125), so we score unaided success per opportunity; the best system answers 75.8% of parts. Accuracy mixes respondent strength with question difficulty; IRT separates the two (Section 4.1).

machine answers rather than 1,485; strict treats both groups alike, so we report both.

Bonus parts are neither interruptible nor pyramidal, so we fit Equation 1 with each part as an item, counting a human answer only if the team gave it without consulting its AI teammates.

4.2 Question Characteristics

Adversarialness estimates show which slots and parts are harder for machines but not what they share, so we describe each with 14 characteristics in four groups (Figure 5; definitions in Appendix A.2). We illustrate each group with the *yellow* tossup (Figure 1). **Visual changes and layout** describe what an image reveals, hides, or rearranges: here, a *visible mask* covers the requested color, so the color is a *hidden target*. **Task demands** describe the symbolic operations and references between works needed to connect the image to an answer, relating to rebus reasoning in COLUMBUS (Kraaijveld et al., 2025) and the use of outside knowledge in OK-VQA (Marino et al., 2019). **Image type** describes the form of visual evidence (here, *artwork*), with subcategories overlapping

those of MMMU (Yue et al., 2024). **Text rules** measure text length and absence-related wording (here, *obscured*), which can require identifying what is missing rather than what is visible (Zhang et al., 2025). We developed the categories on three packets and refined them on the other three; GPT-6 Asra checked consistency, added missing categories, and labeled each slot and part, while the two text features follow fixed rules (Appendix A.2).

Testing associations. For each characteristic, we fit a separate regression across slots or parts, $\delta^c = \alpha + \beta z + \gamma^T \mathbf{v} + \epsilon$, where the outcome δ^c is centered adversarialness, z is the characteristic value (binary, or the log word count for text length), and $\mathbf{v}$ controls for packet and position within a question (for tossups, also squared position and total slot count); a positive β means greater machine-relative difficulty. Because tossup respondents have seen every earlier slot, a tossup slot’s z averages the values of all revealed slots, weighting later slots more (Appendix A.3). We cluster confidence intervals by question, and q is the Benjamini–Hochberg adjusted p -value (Appendices A.3–A.8).

4.3 Hidden Targets as the Clearest Marker

Hidden targets (slots asking for content the image conceals) are the only characteristic that passes the multiple-testing check (Figure 5; dense outcome, 1.17 logits, 95% interval [0.56, 1.79], $q = .003$). Visible masks and absence-related wording also have positive estimates, but do not pass the multiple-testing check ($q = .057$). These characteristics often describe the same questions: an image is masked, the clue says part of the image has been obscured, and the answer names what is missing. Image types such as maps or artwork do not reliably distinguish adversarial examples.

Systems name what is visible. The wrong answers suggest why hidden targets are difficult for machines. At both hidden-target slots (Figures 1 and 9), no AI system answers correctly, and many name an element the image still shows: 10 of 14 answer *red*, the visible color, and 4 of 14 answer *flagpole*. Humans answer correctly at these slots in three and two of five games, so the visible scene and text suffice for some players to infer what is hidden; the systems instead answer about only what is visible. Appendix A.9 gives two related failures.

The association survives most checks but rests on seven tossups. The hidden-target estimate

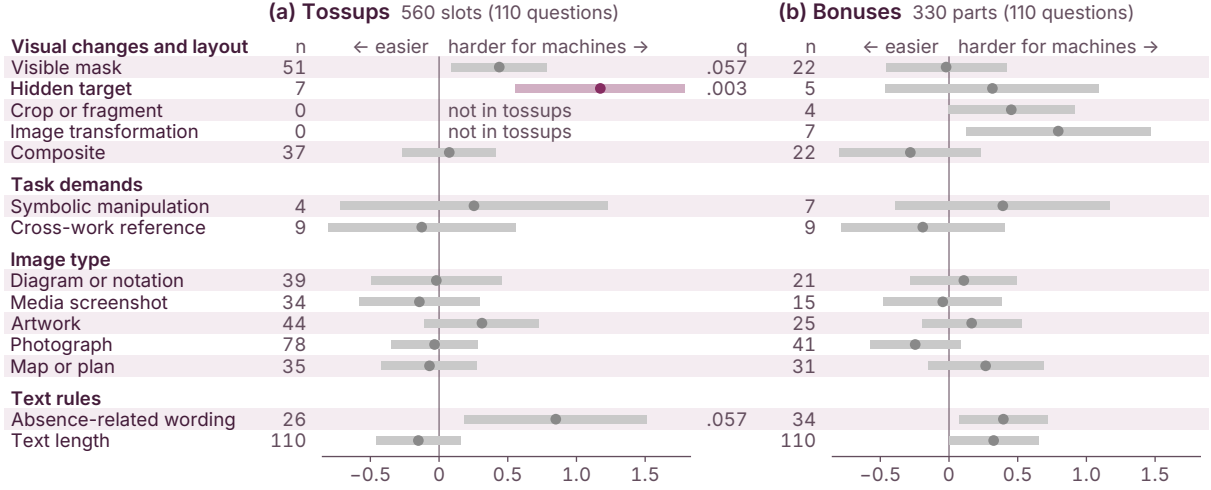


Figure 5: Hidden targets are the only characteristic whose association with adversarialness passes the multiple-testing check ($q = .003$), and only in tossups (a); no bonus characteristic passes (b). Dots are coefficients β in logits (tossups: dense outcome; bonuses: unaided human answers), bars are 95% question-clustered intervals (gray where $q \geq .05$), n counts questions with the characteristic, and q is shown when below .10 (others $\geq .14$). In these data, what a slot asks for, rather than the kind of image it shows, marks adversarial tossups.

stays positive under the strict outcome, where it narrowly misses the check ($q = .054$); with buzz timing modeled separately; with the calibration score of GRACE (Sung et al., 2025a), which does not adjust for skill, as the outcome; with author fixed effects or any one author or question dropped; and when the difficulty model is refitted on resampled questions (Appendices A.5–A.7 and B). It rests, however, on seven tossups from three authors (Table 5): excluding both examples above and adding controls leaves an interval that reaches zero, a within-question comparison shrinks it to 0.14, and no bonus characteristic shows it. In answer to RQ1, hidden targets, in tossups only, are the clearest marker of greater machine-relative difficulty; this comparison of a few questions does not show that masking makes a question adversarial.

5 Adversarial Authoring Strategies

While Section 4 analyzes finalized adversarial questions, we now examine how authors revise images and how revisions reshape the evidence (Section 5.1), then why the model fails even with answer-relevant clues (Section 5.2). Here, a *clue* is any visual evidence pointing to the answer, not only a slot’s text clue. The logs only record the authoring-time model (Qwen3-VL), so a correct-to-wrong change shows that a revision defeats that model, not that the question is adversarial in the sense of Section 4.1, which relies on many models and human performance.

5.1 Revision Categories and Clue-Level Strategies

We link interface events, saved images, and model runs to reconstruct 442 probe–edit–probe pairs (246 tossup and 196 bonus pairs), each comparing the model’s response before and after an author revision (Appendix C). The revision turns a correct response wrong in 80 pairs (CW), leaves a correct response correct in 147 (CC), turns a wrong response correct in 59 (WC), and leaves a wrong response wrong in 156 (WW). We trace each revision at two levels: what changes in the image, and how that change reshapes the route to the answer.

Observable revisions. From model-assisted descriptions of each image pair, refined by hand, we derive five categories and assign one to each pair: **Mask or restrict evidence** hides a label, object, or region (Figures 7a and 14b). **Re-express the same clue** changes the example, view, or representation of a retained clue (Figure 7c). **Switch the identifying clue** substitutes a different clue to the target, such as replacing blood cells with a sickle (Figure 7b). **Resize, rotate, or restyle** changes the scale, layout, orientation, or rendering of existing content. **Reveal or supplement evidence** restores hidden content or adds guidance.

Clue-level strategies. Beyond visual changes, a revision can change how respondents must use the remaining evidence. We develop five strategies from close readings of correct-to-wrong cases and

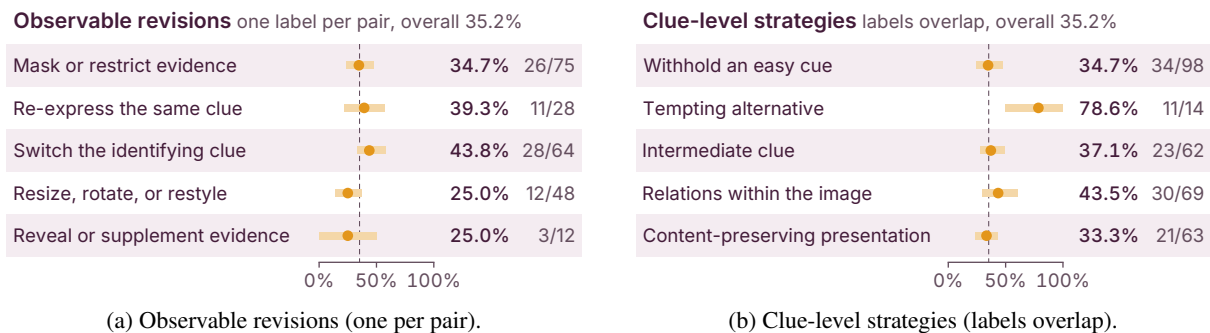


Figure 6: Revisions turn 35.2% of the authoring-time model’s correct answers wrong (dashed line), and strategies that keep the evidence but change how it must be used (b) do so about as often as withholding an easy cue. Marks show $CW/(CC + CW)$, the share of initially correct tests that become wrong, with counts at right and 95% session-bootstrap intervals; clue-level labels include conditional interpretations and restorations, and the rates describe the coded corpus, not causal effects of strategies. Authors need not remove evidence to defeat the model.

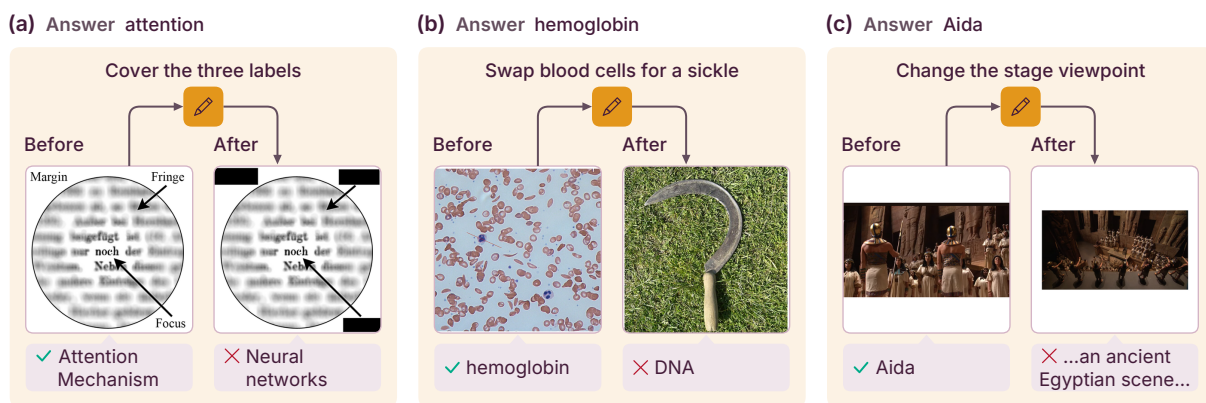


Figure 7: Three revisions, from three observable categories, that turn the authoring-time model’s correct answer wrong. (a) Masking the explanatory labels leaves a focus-and-periphery analogy for attention; the model answers *Neural networks*. (b) Replacing blood cells with a sickle leaves an indirect link to hemoglobin; the model answers *DNA*. (c) A wider stage view keeps the Egyptian theatrical imagery, but the model describes the scene instead of naming *Aida* (answer excerpted). Answers and grades are recorded outputs. In each, a route to the answer plausibly survives the edit, but the model no longer takes it.

apply them to every pair; a pair can receive several:³ (1) **withhold an easy cue**: remove an identifying name, face, or label but leave other evidence that supports the answer; (2) **tempting alternative**: leave a familiar answer that shares features with the target, along with details that distinguish the target; (3) **intermediate clue**: show an element the respondent must connect to the target through background knowledge; (4) **relations within the image**: make the answer depend on how visible elements connect or combine, not on recognizing each one alone; (5) **content-preserving presentation**: change scale, orientation, or style while keeping the identifying content. The two levels do not map one-to-one (Figure 13).

Revisions need not remove clues to defeat the model. Strategies that keep evidence but change how it must be used, namely relations within the image (43.5%), intermediate clues (37.1%), and content-preserving presentation (33.3%), turn initially correct answers wrong about as often as withholding an easy cue (34.7%; Figure 6). Masking or restricting evidence (34.7%) is close to the overall 35.2%, consistent with Section 4.3, where visible masks do not pass the multiple-testing check: a mask need not hide the requested target. Tempting alternatives reach 11 of 14, but that label is easier to assign once the model has produced a distractor. These labels are interpretations, 240 of the 442 conditional, and one session owner contributes most relation labels (Appendix C.2).

³Tables 8 and 9 give the coding boundaries; Appendix C.2 explains their development, and Appendix C.4 gives examples.

5.2 Model Failures Despite Surviving Clues

To see why the model fails when answer-relevant clues survive an edit, we read all 80 correct-to-wrong pairs. In our judgment, a route to the answer survives in 27 concretely and in 44 conditionally (Appendix C.4). The model’s explanations show three overlapping patterns, in 42, 43, and 15 pairs; 12 pairs show none.

A familiar guess overrides stronger evidence.

The model can favor a familiar answer over evidence it explicitly recognizes. After an author masks location labels on a ski map, the model places a name in Italy but answers Austria based on German-language place names (Figure 14a). Similarly, it recognizes spreadsheet-specific controls in an edited Google Sheets toolbar but answers Google Docs based on a shared font dropdown (Figure 14b). These cases suggest that a human advantage may lie not in additional knowledge, but in weighing which clues are more diagnostic.

Surviving clues go unconnected. The model can recognize each element yet miss what the elements mean together or what they point to through background knowledge. In Figure 14c, fruit and exhaled breath together suggest fruity breath, a symptom of diabetic ketoacidosis, but the model reads them separately, as healthy eating and environment, and answers diabetic retinopathy; the evidence remains for a knowledgeable human to connect.

Recognition changes with presentation. Identifying content can survive an edit while recognition does not. In Figure 7c, a wider view of an Egyptian stage scene shifts the model from *Aida* to a generic description of the scene, although the clues to the opera remain.

In answer to RQ2, authors defeat the authoring-time model not only by withholding easy cues but also by changing how surviving evidence must be weighed, connected, or recognized; whether humans can still answer the revised questions remains untested.

6 Related Work

Adversarial collection and human-grounded evaluation.

Prior adversarial VQA collection asks people to write questions about supplied images (Sheng et al., 2021; Li et al., 2021); our authors also replace and edit the images, and the logs keep revisions that fail. Because the model used

during collection shapes the data (Bartolo et al., 2020), we separate authoring-time feedback from later gameplay, and our before/after pairs resemble contrast sets and counterfactually augmented text (Gardner et al., 2020; Kaushik et al., 2020) but come from model-guided revisions. IRT separates respondent ability from item difficulty in NLP evaluation (Rodriguez et al., 2021), CAIMIRA characterizes complementary human and machine abilities (Gor et al., 2024), and AdvScore grounds adversarialness in human ability (Sung et al., 2025b); we extend AdvScore to cumulative clues and selectively observed gameplay.

Visual grounding and counterfactual evaluation.

Counterfactual edits that remove or mask image regions test whether VQA models use the relevant evidence (Agarwal et al., 2020); counterfactual samples that mask critical objects or words reduce language bias in training (Chen et al., 2020); minimal counterfactual images that change a model’s answer support interpretation (Boukhers et al., 2022); and focused question augmentations measure whether answers stay consistent (Rosenberg et al., 2021). Our authors make such edits to create questions instead, and we distinguish recognizing visible content from recovering a hidden target. Authors’ revisions are naturally occurring counterfactuals chosen to defeat a model; controlled versions of our hidden-target edits, with masked, unmasked, and non-target masked images, would connect the two. Appendix E discusses modality shortcuts (Li et al., 2024; Uebayashi et al., 2026) and adaptive evaluation.

7 Conclusion

In human–AI quizbowl play, tossup slots that ask for what an image hides are the clearest marker of machine-relative difficulty, more than image type, although this association rests on seven tossups from three authors. Authors made questions harder for the authoring-time model not only by withholding easy cues, but also by changing how surviving evidence must be prioritized, connected, or recognized. Together, these findings suggest that adversarialness depends less on what visual evidence is available than on what a question requires respondents to do with it, and that expert authors with a model in the loop, failed edits included, are a productive way to find where VLMS lag skilled humans. Next steps are controlled versions of these edits with human answers to the revised questions,

and using such questions to evaluate and train models that reason about absent evidence.

Limitations

Questions and participants. Our findings rest on 240 English quizbowl questions from a small group of authors (Table 1), many of them experienced quizbowl writers, and answered by skilled quizbowl players and by the AI systems submitted to one competition. Quizbowl rewards specialized knowledge of a largely North American trivia canon, so the questions that are adversarial here may differ from those that are adversarial in everyday VQA, in other languages and cultures, or for less experienced people. Authors probed mainly one model (Qwen3-VL) while writing, and questions that defeat it need not defeat the systems that later answered them. Skilled quizbowl players are not a universal standard of human answerability: the collection suits studying gaps between machines and expert humans and how adversarial questions are written, not everyday visual assistance or the general population.

Measuring adversarialness. Human first buzzes are sparse and selectively observed; bonus unaided success also reflects consultation decisions. Our estimates are relative: they rank questions within each format by machine-relative difficulty, but do not measure an absolute human advantage. Regression intervals condition on estimated difficulty; refitting on resampled questions gives a similar interval (Appendix A.7), but uncertainty in the annotations remains unmodeled. Partial-answer policies can affect apparent failures after successful recognition, and bonus prompt/image-version identity has not been independently verified.

Labels. We used one LLM-based assistant for RQ1 content coding and answer-alias review, with outcomes available during review, so independent human validation remains necessary. In particular, the hidden-target result rests on a few tossups and authors (Section 4.3; identified through session usernames), making author, topic, and template effects difficult to separate. Adjusting the listed tests does not resolve adaptive feature selection. For RQ2, two models produced the initial image descriptions (Appendix C.2); the resulting revision and strategy codes likewise lack independent human validation and inter-rater reliability, and the strategies were developed with access to model

failures.

Authoring logs. RQ2 observes adaptive probe histories rather than randomized edits, and its 442 pairs come from 114 sessions owned by 17 of the 27 authors. Its outcomes are responses of the authoring-time model only; we do not measure whether humans can still answer the revised questions. Image identity and stable stored keys do not establish semantic continuity, and complete test-time wording (clue text was not versioned), text-only tests, and historical human responses are unavailable. Generated explanations may rationalize an answer rather than reveal the model’s reasoning. No text-only or image-only ablation, or controlled editing intervention, establishes a causal or uniquely visual mechanism. The analysis therefore identifies candidate construction patterns and local feedback associations, not a causal ranking of strategies that preserve human answerability. These constraints limit the findings to evidence-supported hypotheses rather than a universal taxonomy of adversarial VQA.

Ethics Statement

Participants. Question authors were recruited through the competition’s public call for writers, were paid \$5 per accepted question, could win writing prizes (\$50–\$100 for the best and the most adversarial question and packet), and signed an IRB-approved consent form. All players signed an IRB-approved consent form before playing, in both the in-person and the online games. Human game-play data are buzzes and answers from competition games; buzzes are logged per player, but our analysis pools players and reports no player-level results. We exclude the developer’s and organizer’s main accounts and one unreviewed bulk-entry account from authoring statistics; author and submission counts also exclude four further developer and test accounts.

Images. Authors chose and edited images for a research competition; many depict copyrighted works such as film stills, artwork, and book covers, which we use for research purposes.

Risks. The editing strategies we document, such as masking question content, could also be used to craft inputs that mislead deployed VLMS; we report them as aggregate patterns that also expose current model weaknesses. Some images show identifiable people, for example in historical photographs, so

reusing them raises concerns about likeness as well as copyright. Authoring logs could help identify authors through their editing histories and topics; we therefore report authoring behavior only in aggregate or under pseudonymous identifiers.

Use of AI assistants. We used LLM-based assistants to check and extend the characteristic categories, label question characteristics (Section 4.2), review answer aliases, and describe image revisions and coded revision categories and strategies (Section 5); we treat these labels as exploratory (see Limitations). We also used AI assistants to edit text, produce figures and screenshots, and cross-check reported statistics against the logged data; we reviewed all changes.

References

- Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. [Towards causal VQA: Revealing and reducing spurious correlations by invariant and covariant semantic editing](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. [Don’t just assume; look and answer: Overcoming priors for visual question answering](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. [VQA: Visual question answering](#). In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. [Qwen3-VL technical report](#). *Preprint*, arXiv:2511.21631.
- Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. [Beat the AI: Investigating adversarial human annotation for reading comprehension](#). *Transactions of the Association for Computational Linguistics*, 8:662–678.
- Yoav Benjamini and Yosef Hochberg. 1995. [Controlling the false discovery rate: A practical and powerful approach to multiple testing](#). *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300.
- Zeyd Boukhers, Timo Hartmann, and Jan Jürjens. 2022. [COIN: Counterfactual image generation for visual question answering interpretation](#). *Sensors*, 22(6):2245.
- Samuel R. Bowman and George E. Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855. Association for Computational Linguistics.
- Jordan Boyd-Graber and Benjamin Börschinger. 2020. What question answering can learn from trivia nerds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7422–7435. Association for Computational Linguistics.
- A. Colin Cameron and Douglas L. Miller. 2015. [A practitioner’s guide to cluster-robust inference](#). *Journal of Human Resources*, 50(2):317–372.
- Stuart K. Card, George G. Robertson, and Jock D. Mackinlay. 1991. [The information visualizer, an information workspace](#). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Reaching through Technology (CHI ’91)*, pages 181–186, New Orleans, Louisiana, United States. ACM Press.
- Centers for Disease Control and Prevention. 2024. [Diabetic ketoacidosis](#). Accessed October 10, 2026.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. [Are we on the right way for evaluating large vision-language models?](#) In *Advances in Neural Information Processing Systems*, volume 37, pages 27056–27087. Curran Associates, Inc.
- Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. [Counterfactual samples synthesizing for robust visual question answering](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Paul De Boeck and Ivailo Partchev. 2012. [IRTrees: Tree-based item response models of the GLMM family](#). *Journal of Statistical Software*, 48.
- Fabrizio Dell’Acqua, Edward McFowland, Ethan Mollick, Hila Lifshitz, Katherine C. Kellogg, Saran Rajendran, Lisa Kraye, François Candelon, and Karim R. Lakhani. 2026. [Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality](#). *Organization Science*, 37(2):403–423.
- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, and 1 others. 2020. Evaluating models’ local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP*

- 2020, pages 1307–1323. Association for Computational Linguistics.
- Maharshi Gor, Hal Daumé III, Tianyi Zhou, and Jordan Boyd-Graber. 2024. [Do great minds think alike? Investigating human-AI complementarity in question answering with CAIMIRA](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21533–21564, Miami, Florida, USA. Association for Computational Linguistics.
- Maharshi Gor, Yoo Yeon Sung, Yu Hou, Eve Fleisig, Zhu Irene Ying, Tianyi Zhou, and Jordan Lee Boyd-Graber. 2026. [AI, take the wheel: What drives delegation and trust in human-computer cooperative question answering?](#) In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 8664–8689, San Diego, California, United States. Association for Computational Linguistics.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. [Making the V in VQA matter: Elevating the role of image understanding in visual question answering](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.
- Valentin Hofmann, David Heineman, Ian Magnusson, Kyle Lo, Jesse Dodge, Maarten Sap, Pang Wei Koh, Chun Wang, Hannaneh Hajishirzi, and Noah A. Smith. 2025. [Fluid language model benchmarking](#). In *Second Conference on Language Modeling*.
- Paul W. Holland and Howard Wainer, editors. 1993. *Differential Item Functioning*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Ken Jennings. 2006. *Brainiac: Adventures in the curious, competitive, compulsive world of trivia buffs*. Villard.
- Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.
- John D. Kalbfleisch and Ross L. Prentice. 2002. *The Statistical Analysis of Failure Time Data*. Wiley.
- Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. [Dynabench: Rethinking benchmarking in NLP](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.
- Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2023. [\(QA\)²: Question answering with questionable assumptions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8466–8487, Toronto, Canada. Association for Computational Linguistics.
- Koen Kraaijveld, Yifan Jiang, Kaixin Ma, and Filip Ilievski. 2025. [COLUMBUS: Evaluating COgnitive Lateral Understanding through Multiple-choice re-BUSes](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):4410–4418.
- Zixuan Lan, Luzhe Sun, Matthew R. Walter, and Jiawei Zhou. 2026. [Seeing without looking: Do vision-language benchmarks really test vision?](#) *Preprint*, arXiv:2605.22903.
- Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. 2024. [NaturalBench: Evaluating vision-language models on natural adversarial samples](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 17044–17068. Curran Associates, Inc.
- Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. 2021. [Adversarial VQA: A new benchmark for evaluating the robustness of VQA models](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2042–2051.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. [OK-VQA: A visual question answering benchmark requiring external knowledge](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jordan Meyer, Nick Padgett, Cullen Miller, and Laura Exline. 2024. [Public domain 12M: A highly aesthetic image-text dataset with novel governance mechanisms](#). *Preprint*, arXiv:2410.23144.
- Robert B. Miller. 1968. [Response time in man-computer conversational transactions](#). In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I (AFIPS '68 (Fall, part I))*, San Francisco, California. ACM Press.

- Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. [AmbigQA: Answering ambiguous open-domain questions](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. [Adversarial NLI: A new benchmark for natural language understanding](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.
- Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.
- Pedro Rodriguez, Joe Barrow, Alexander Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. [Evaluation examples are not equally informative: How should that change NLP leaderboards?](#) In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.
- Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan L. Boyd-Graber. 2019. [Quizowl: The case for incremental question answering](#). *CoRR*, abs/1904.04792.
- Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. [QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension](#). *ACM Computing Surveys*, 55(10):1–45.
- Daniel Rosenberg, Itai Gat, Amir Feder, and Roi Reichart. 2021. [Are VQA systems RAD? Measuring robustness to augmented data with focused interventions](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 61–70, Online. Association for Computational Linguistics.
- Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. 2021. [Human-adversarial visual question answering](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 20346–20359. Curran Associates, Inc.
- Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. [WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449. ACM.
- Elias Stengel-Eskin, Jimena Guallar-Blasco, Yi Zhou, and Benjamin Van Durme. 2023. [Why did the chicken cross the road? Rephrasing and analyzing ambiguous questions in VQA](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10220–10237, Toronto, Canada. Association for Computational Linguistics.
- Yoo Yeon Sung, Eve Fleisig, Yu Hou, Ishan Upadhyay, and Jordan Lee Boyd-Graber. 2025a. [GRACE: A granular benchmark for evaluating model calibration against human calibration](#). *Preprint*, arXiv:2502.19684.
- Yoo Yeon Sung, Maharshi Gor, Eve Fleisig, Ishani Mondal, and Jordan Boyd-Graber. 2025b. [Is your benchmark truly adversarial? AdvScore: Evaluating human-grounded adversarialness](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 623–642, Albuquerque, New Mexico. Association for Computational Linguistics.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. [Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578.
- Shunki Uebayashi, Kento Masui, Kyohei Atarashi, Han Bao, Hisashi Kashima, Naoto Inoue, Mayu Otani, and Koh Takeuchi. 2026. [Evaluating cross-modal reasoning ability and problem characteristics with multimodal item response theory](#). In *International Conference on Learning Representations*, volume 2026, pages 97794–97817.
- Wim J van der Linden and Ronald K Hambleton. 2013. *Handbook of modern item response theory*. Springer Science & Business Media.
- Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.
- Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. [CREPE: Open-domain question answering with false presuppositions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480, Toronto, Canada. Association for Computational Linguistics.
- 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, and 3 others. 2024. [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 (CVPR)*, pages 9556–9567.

Yuhui Zhang, Yuchang Su, Yiming Liu, and Serena Yeung-Levy. 2025. [NegVQA: Can vision language models understand negation?](#) In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3707–3716, Vienna, Austria. Association for Computational Linguistics.

	Tossups	Bonuses
Authoring: 27 authors ^a , 481 sessions		
Submitted	149	127
Played in the six packets	120	120
Analysis set		
Questions	110	110
Slots or parts	560	330
AI systems ^b	14	10

Table 1: Editors accepted 245 of the 276 submitted questions. The analysis set keeps questions with compatible human and AI records (Appendix A.1). ^aTwo authors wrote jointly under another author’s account.

A RQ1 Analysis Details

A.1 Data Coverage and Difficulty Estimation

Coverage. The six tournament packets contain 120 tossups (610 slots) and 120 bonuses (360 parts). We inspect all 970 units; compatible gameplay supports estimates for 560 slots from 110 tossups and 330 parts from 110 bonuses (Table 1). Unobserved or incompatible units remain in the content inventory but receive no DIF estimate. The fitted tossup set contains 490 human episodes (334 first buzzes, of which 256 are correct, and 156 censored episodes) and 7,558 machine slot-boundary observations from 14 systems. An episode is one tossup in one game; its event is the first buzz by a human player on either team, and buzzes by AI teammates never count as human events. In 153 of the 156 censored episodes, an AI teammate answered correctly before any human buzzed; the other three reach the end of the question after only incorrect AI buzzes. In 34 episodes an AI had buzzed incorrectly by the end of the episode; the model does not represent that such a buzz locks its team out of the tossup. We exclude 14 further human episodes: for seven packet-6 tossups, the token count recorded in the two in-person games differs from the packet text, and these tossups remain estimable from the other games. We also exclude six episodes on two tossups outside the 240 questions, which three games read in place of two packet-6 tossups.

Machine answers. Machine evidence comes from the systems’ saved offline shared-task answers, recorded at checkpoints a median of six tokens apart. Twelve systems answered all 110 fitted tossups and two answered 90 (each lacks one packet), giving 7,643 system–slot pairs; we drop 85 pairs without an answer exactly at the slot’s end, 79 of them from one system whose answers stop early or skip slot ends. Of the 560 slots, 549 have

13 or 14 machine answers. Humans and systems receive the same clue text and slot or part images; a bonus part’s optional reference image, which is not needed for the answer, is shown only to players (Appendix D).

Tossup likelihood. For each first-buzz opportunity, we record the preceding non-buzzes, the first buzz, and whether the answer is correct. Following the response-tree framework of De Boeck and Partchev (2012), each slot is a branch between waiting and buzzing, followed by a correct or incorrect branch after a buzz. Writing the buzz probability in Equation 2 as h_{ris} and the correctness probability conditional on buzzing as $\pi_{ris} = \sigma(\theta_r - b_{is}^{gr})$, a correct buzz contributes $h_{ris}\pi_{ris}$ to the likelihood, an incorrect buzz contributes $h_{ris}(1 - \pi_{ris})$, and a non-buzz contributes $1 - h_{ris}$. These contributions jointly inform difficulty because both probabilities depend on b_{is}^{gr} . For example, waiting through slot 1 and buzzing correctly in slot 2 contributes $(1 - h_{ri1})h_{ri2}\pi_{ri2}$; the slot-2 difficulty b_{i2}^{gr} appears in both h_{ri2} and π_{ri2} . If the opportunity ends before a buzz, we use only the observed non-buzzes; the unobserved answer does not count as incorrect (Kalbfleisch and Prentice, 2002). Under the dense outcome, machine correctness is instead observed at every available slot boundary.

Bonuses. The bonus model uses 1,503 human opportunities and 2,850 machine responses over 330 shared parts: five systems answered all 330 parts and five answered only the 240 parts in packets 1–4, so each part has ten or five machine answers. Human opportunities comprise 766 unaided choices, 612 consultations, and 125 abstentions. Bonuses are joined by question and part identifiers; this does not independently verify prompt/image identity.

Grading. Machine correctness keeps the grades recorded in the systems’ outputs and also accepts an answer that, after normalization, exactly matches an accepted answer in the packet or one of 312 alias entries added for 38 tossup and 80 bonus answers after an LLM-assisted review (see Limitations). A matching token alone earns no added credit; we do not re-review original acceptances, 1,148 of 17,050 of which match no accepted answer or alias exactly. This repair changes 4,688 of 35,703 machine responses from incorrect to correct, including 1,108 of the 7,558 dense tossup observations (in 31 of 110 tossups) and 263 of the 2,850 analyzed bonus responses.

Estimation. We fit tossups and bonuses separately; their centered scores do not share an externally calibrated origin. The implementation uses the equivalent midpoint parameterization $\tilde{b}_i = b_i + \delta_i/2$, with human difficulty $\tilde{b}_i - \delta_i/2$ and machine difficulty $\tilde{b}_i + \delta_i/2$. Tossup maximum a posteriori estimation uses zero-mean Gaussian priors with standard deviations 1.5 for ability, 2.5 for midpoint difficulty, and 1.5 for DIF. Adjacent-slot increments have standard deviations 1.0 for midpoint difficulty and 0.75 for DIF. Buzz thresholds have mean 1 and standard deviation 2; the strict fit includes a threshold per system. All human tossup episodes share one pooled ability and one buzz threshold, so for humans θ_r and c_r are population-level parameters rather than per-player parameters; the bonus model instead gives humans one ability per game room (five rooms) and each system its own. All six primary and sensitivity fits (dense, strict, both generous-answer variants, and both bonus outcomes) converge with maximum absolute gradient below .005.

A.2 Characteristics and Coding Boundaries

Table 2 lists the 14 characteristics; below we give their coding boundaries, examples, and distinctions between related labels.

Visual changes and layout. These five characteristics are binary and describe what an image reveals, hides, or rearranges.

Visible mask records an opaque mask or redaction anywhere in the image. It does not specify whether the covered region contains the answer, a caption, or another clue. *Hidden target* records whether the queried object, person, or color is covered or explicitly removed. This distinction depends on the question and accepted answer as well as the image. For example, the concealed yellow region in Figure 9 (left) is the target because the question asks for that color. In the Jamaica tossup (Figure 2, slots 3 and 4), masks cover words inside the rebuses; the requested country is not itself a concealed visual target. A visible mask can therefore receive a positive masking label without a positive hidden-target label. The definition also admits explicit removal without an opaque mask, but every hidden-target unit in our data (27 tossup slots and 10 bonus parts) also has a visible mask.

Crop or fragment marks only a stated or implied fragment task; we do not infer crops from an image’s editing history. *Composite* describes

the layout within a single image. These two labels distinguish the availability and arrangement of evidence within an image.

Task demands. These two characteristics are assigned at parent-question level and carried over to its slots or parts, recording the shared task even when an individual image does not show every step.

Symbolic manipulation covers explicit rebuses (visual word puzzles), anagrams, and operations on names, letters, or codes, including zip codes. For example, the Peru bonus requires turning the pictured word *mail* into the capital *Lima*, then naming its country (Figure 11b). Our label also includes other word and letter operations; it records the task demand rather than coding every step separately.

Cross-work reference marks identifying a source through a reference, adaptation, or impersonation in another work. Both labels describe what the question requires, independent of model behavior.

Image type. These five binary labels are not mutually exclusive; labels may overlap when an image contains several kinds of material. Further, none of these five labels applies to the 21 of 970 units with no image or whose only image is a text document.

Text rules. Both features are computed automatically from the saved text.

Absence-related wording is a binary indicator obtained through case-insensitive keyword matching with word boundaries. The vocabulary covers *not*, *except*, *instead*, *without*, *missing*, *removed*, *obscured*, *redacted*, and *black(ed) out*. It includes both explicit exclusion and references to hidden content. A question can use *not* without any image editing, while a hidden target can occur without any listed keyword. The indicator therefore records wording separately from the visual mask and hidden-target labels. A match does not identify the word’s scope or establish which reasoning operation is required.

Text length is $\log(1 + w)$, where w counts tokens separated by whitespace in the current slot or bonus part. Longer text can supply additional clues or introduce more constraints to track, so length does not necessarily imply greater difficulty. Bonus text includes the shared lead-in; its visual labels describe the focal part image rather than the separate lead-in example. We test text length as its own characteristic, and we do not control for text length when fitting regressions for other characteristics.

Group	Characteristic	What is recorded
Visual changes and layout	Visible mask	An opaque mask or redaction is visible.
	Hidden target	The queried object, person, or color is covered or removed.
	Crop or fragment	The task explicitly uses a cropped image or fragment.
	Image transformation	Explicit changes such as rotation, redrawing, or parody, excluding masking.
	Composite	Multiple panels or objects are combined in the focal image.
Task demands	Symbolic manipulation	Rebus, anagram, or explicit manipulation of names, letters, or codes.
	Cross-work reference	Identify a source through a reference or adaptation in another work.
Image type	Diagram or notation	Diagram, chart, equation, notation, or emblem.
	Media screenshot	Still from a film, television show, video, or game.
	Artwork	Artwork, illustration, comic, or cover.
	Photograph	Photograph, including photographs of sculptures or buildings.
	Map or plan	Map, satellite image, floor plan, or site plan.
Text rules	Absence-related wording	Contains a keyword such as <i>not</i> , <i>without</i> , or <i>obscured</i> .
	Text length	$\log(1 + \text{word count})$; bonus text includes the lead-in.

Table 2: The 14 VQA characteristics tested for association with adversarialness, in four groups. Appendix A.2 gives coding boundaries and examples.

A.3 Association Models and Statistical Adjustments

Tossup respondents see all previously revealed clues, so we measure cumulative exposure to each characteristic using a weighted average that favors later slots:

$$x'_s = \frac{\sum_{u=1}^s u x_u}{\sum_{u=1}^s u}, \quad (3)$$

where x_u is the original value of a characteristic in slot u . For example, if a hidden target appears in the first slot but not the second, its value at slot two is $1/3$; if it appears only in the second slot, the value is $2/3$. Parent-level task labels remain constant under this averaging. For text length, we average the per-slot log word counts; we do not take the log of the total words revealed. For bonuses, each part uses its own visual and text features together with the shared parent-level task labels.

Relative slot position is $(s - 1)/(S_i - 1)$, with zero for a single-slot question. All regressions include an intercept and packet indicators, plus relative slot position, its square, and total slot count for tossups, and part position for bonuses.

We cluster standard errors by question (Cameron and Miller, 2015) so that repeated observations from one question do not make an association appear more certain than it is. We also adjust for testing many characteristics using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995), using adjusted $q < .05$ as the threshold for statistical evidence.

We use a finite-sample corrected cluster sandwich covariance and t intervals with $G - 1$ degrees

of freedom for G source-question clusters. We apply Benjamini–Hochberg adjustments separately within each outcome/specification’s estimable feature family; they do not account for selecting the codebook after inspecting outcomes. Author checks attribute each question to the owner of its authoring session, who may not have made every final edit.

A.4 Dense and Strict First-Buzz Estimates

We compare two ways of using machine responses to estimate tossup adversarialness. The dense model uses answer correctness at every available slot endpoint; the strict model uses first-buzz correctness and censoring records. Both use the same human evidence and the same 560 slots from 110 tossups, with the answer matching used in the main analysis. We center each model’s difficulty differences separately and summarize each question by the token-weighted average of its slot scores, $Q_i = \sum_s w_{is} \delta_{is}^c / \sum_s w_{is}$, where w_{is} is the number of tokens in slot s of tossup i . Table 3 compares the resulting scores, and Figure 8 compares their associations with question characteristics.

Table 3 shows substantial agreement in ordering, but the highest-scoring items depend on the machine response definition. The hidden-target coefficient remains positive under both definitions, although it is smaller under the strict outcome. We therefore use dense scores to study answer content and strict scores to check whether associations persist under the systems’ buzzing policies.

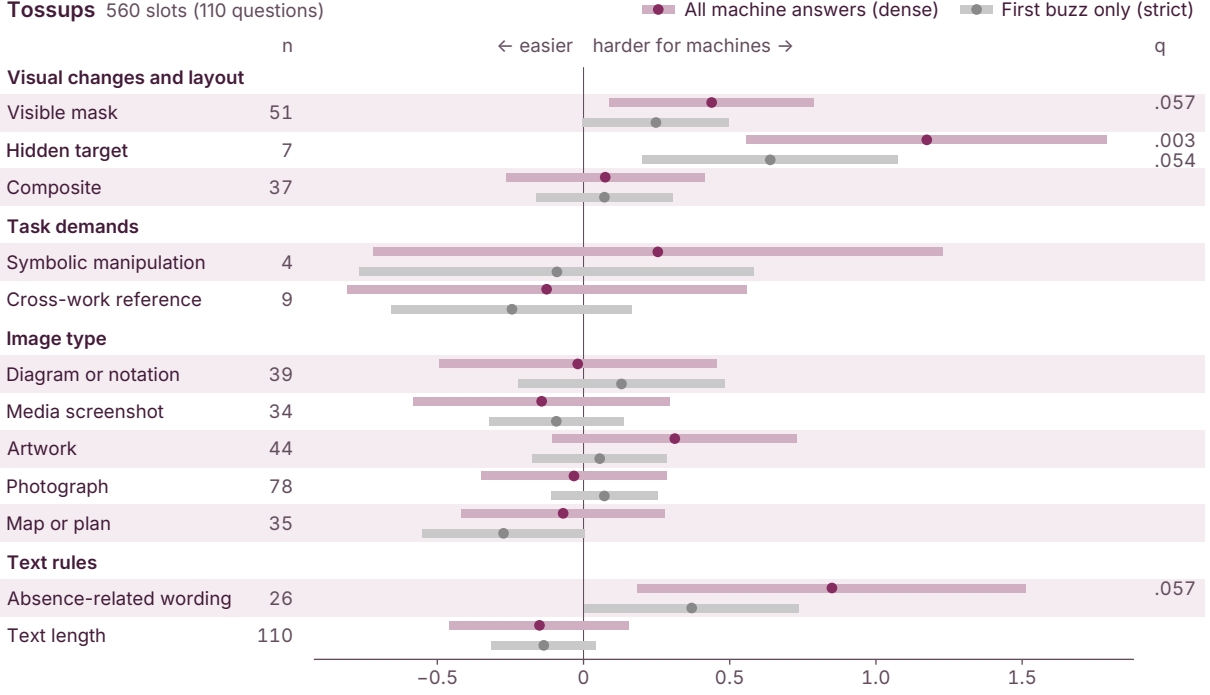


Figure 8: Hidden targets have the largest positive association under both tossup outcomes, but under the strict outcome the estimate is smaller (0.64, $q = .054$) and no characteristic passes the multiple-testing check. In each row, the upper bar is dense (all machine answers) and the lower, gray bar is strict (first buzz only); dots are regression coefficients with the main analysis’s position-weighted cumulative characteristics and controls, and bars are 95% question-clustered intervals conditional on fitted DIF. n and q are as in Figure 5; crop or fragment and image transformation have no variation among eligible tossup slots. Counting only first buzzes weakens the hidden-target association but does not reverse it.

Level	N	Spearman	Same sign	Top overlap
Slots	560	.774	78.6%	48.2%
Questions	110	.810	81.8%	54.5%

Table 3: Dense and strict outcomes rank slots and questions similarly but only share about half of their most adversarial units. Sign agreement uses scores centered within each fit, and top overlap is the share of each fit’s highest-scoring 10% that the other fit also places there.

A.5 Sensitivity Analyses

We examine whether the hidden-target association depends on the outcome definition, accompanying characteristics, or a few questions. For bonuses, the primary outcome counts unaided correct answers over all opportunities. The alternative conditions on choosing to answer alone, and its bonus hidden-target estimate is $-0.01 [-0.71, 0.68]$ ($q = .97$), and no bonus characteristic passes the multiple-testing check (all $q \geq .15$). Table 4 reports hidden-target checks under both tossup outcomes. Equal weights give similar primary estimates as position weights (1.15 versus 1.17).

The expanded controls add visible mask, map

or plan, artwork, media screenshot, diagram or notation, composite, and absence-related wording indicators. The local comparison instead uses question fixed effects and the current slot’s feature; its null result limits the interpretation to differences between question constructions. The generous policy also accepts partial answers on three hidden-target tossups (Table 5): *flag* for American flags, *timer* for sand timer, and *Peter Rabbit* for rabbits; it is a deliberately generous check, not a proposed answer policy. Author fixed effects and dropping one hidden-target question or author at a time (Table 4) refit only the regression, with DIF scores held fixed.

A.6 Commitment versus Correctness

We test a more flexible slot-wise model while retaining the same dense machine answers, 560 slots, human episodes, grades, and characteristic regressions used in Figure 5. Instead of sharing correctness difficulty with the human buzz hazard, we estimate an independent commitment propensity λ_H and slot threshold a_{is}^H :

$$P(B_{is} = 1 \mid B_{i,<s} = 0) = \sigma(\lambda_H - a_{is}^H). \quad (4)$$

Hidden-target comparison	Dense estimate [95% interval]	Strict estimate [95% interval]
Primary position-weighted feature	1.17 [0.56, 1.79]	0.64 [0.20, 1.07]
Equal-weight cumulative feature	1.15 [0.53, 1.77]	0.62 [0.20, 1.05]
Additional visual/text controls	1.07 [0.31, 1.83]	0.60 [0.07, 1.14]
Exclude yellow and American flags	0.69 [0.42, 0.97]	0.41 [-0.04, 0.87]
Exclude both, additional controls	0.53 [-0.01, 1.06]	0.34 [-0.24, 0.92]
Generous partial-answer policy	0.89 [0.07, 1.71]	0.56 [0.08, 1.04]
Author fixed effects	1.20 [0.57, 1.83]	0.63 [0.18, 1.08]
Drop one hidden-target question (7 fits)	0.97 to 1.38 (lower $\geq$ 0.37)	0.50 to 0.81 (lower $\geq$ 0.03)
Drop one hidden-target author (3 fits)	0.68 to 1.70 (lower $\geq$ 0.40)	0.57 to 0.70 (lower $\geq$ -0.01)
Local feature, question fixed effects	0.14 [-0.17, 0.45]	-0.08 [-0.29, 0.14]
Any hidden-target question (mean DIF)	1.04 [0.48, 1.60]	0.52 [0.17, 0.87]

Table 4: The dense hidden-target estimate stays positive in every check, but its interval reaches zero when the two strongest questions are excluded and controls are added, and within questions (local feature) it nearly vanishes. The primary row uses position-weighted cumulative characteristics; the next row uses equal weights. All intervals condition on fitted DIF. Drop-one rows give the range of estimates and the lowest lower bound of their intervals. The final row uses mean slot DIF per question and a binary any-hidden-target indicator with packet and slot-count controls; other rows use slot outcomes. The association separates questions more than slots within a question.

Author	Answer	Hidden slots	Mean centered DIF	
			hidden	other
A	<i>yellow</i>	3/5	2.31	1.80
A	American flags	5/5	1.80	—
A	rhinoceros	3/5	0.79	0.30
A	rabbits	2/5	0.24	-0.19
B	sand timer	7/7	0.37	—
B	Donald Glover	6/7	0.58	-0.10
C	Burj Khalifa	1/4	0.07	1.03

Table 5: In four of the five hidden-target tossups that have other slots, the hidden-target slots are more adversarial than the rest of the question. Across all 560 slots, centered DIF has mean 0 and standard deviation 0.74.

We fit the commitment component to 4,454 slot-level survival observations: 1,529 from the 490 human episodes (one pooled propensity) and 2,925 from the 14 systems’ strict first-buzz records (one propensity per system and a human-machine threshold difference per slot). The systems’ records do not enter the correctness component, whose estimates are compared below. Correctness retains the primary IRT formulation and priors; the independent commitment parameters use the same Gaussian shrinkage and adjacent-slot smoothing structure. Human first buzzes contribute survival, an event, and correctness, whereas censored episodes only contribute survival. The two likelihood components factorize, so waiting no longer informs correctness difficulty through shared parameters.

Figure 10 shows the dense-only comparison. Hidden targets remain the only characteristic passing the multiple-testing check: their coefficient decreases from 1.174 to 0.800 logits (95% question-cluster interval [0.404, 1.196], $q = .0014$). Visible



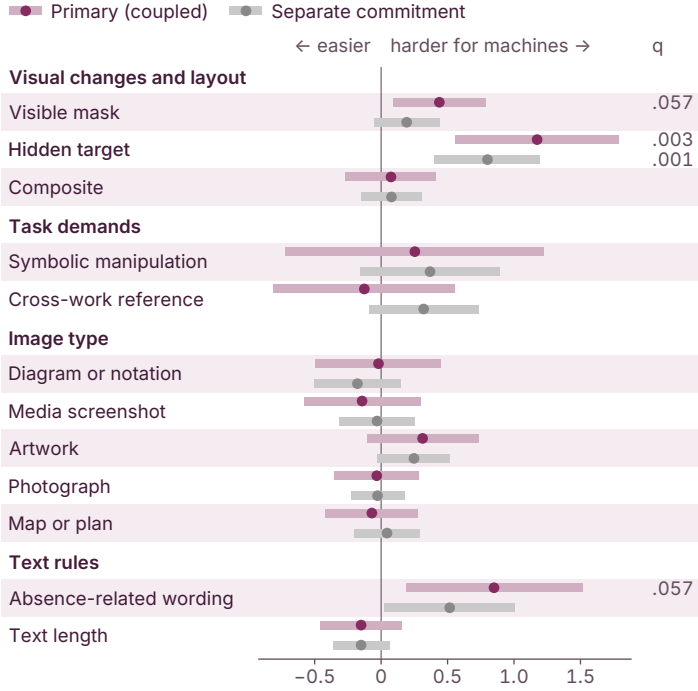
Figure 9: At both hidden-target slots, no AI system answers correctly (0/14), with many of them naming what the image still shows. In contrast, humans answer correctly in three (left) and two (right) of five games. Left: the question of Figure 1; right: a tossup asking what a mask covers in a Dorothea Lange photograph.

masking and absence-related wording have weaker associations; slot ranks correlate at $\rho = .784$. The plotted intervals condition on fitted scores. Separately propagating their full-Hessian Laplace covariance gives a hidden-target coefficient interval of [0.266, 1.335] for this fixed corpus. Thus the qualitative hidden-target finding persists, although score magnitudes and secondary associations change.

A.7 Refitting the Difficulty Model

The association tests above condition on one fit of the difficulty model. We refit the dense model from its saved observation rows, which reproduces the

(a) Characteristic associations



(b) Slot estimates 560 slots



Figure 10: Separating buzz timing from correctness shrinks the hidden-target coefficient from 1.17 to 0.80, but it remains the only characteristic that passes ($q = .001$), and slot estimates stay correlated (Spearman $\rho = .784$). (a) Characteristic coefficients with 95% question-clustered intervals conditional on fitted scores (upper bar: primary model; lower, gray bar: separate commitment); crop and transformation have no variation. (b) Centered slot DIF estimates under the two models; the dashed line marks equality. Both models use the same dense machine answers, labels, and regression controls. The hidden-target finding does not rely on timing sharing difficulty with correctness.

reported slot estimates (maximum absolute difference below 10^{-5}), and rerun the primary hidden-target regression after each change (Table 6). Resampling the 110 tossups with replacement and refitting both stages 200 times carries the uncertainty of the difficulty estimates into the regression. The other rows give each of the five game rooms its own human ability, replace the alias-repaired machine grades with the original grades, and tighten or loosen the prior on the human-machine difference and the smoothing between adjacent slots. Finally, we simulate 50 data sets from the fitted model with the observed structure (the same game rooms, the same points at which an AI teammate answered first, and the same machine observations) and refit each. Recovered slot estimates correlate .74 with the true values on average, and the hidden-target coefficient is positive in every run but averages 0.60 against a true 1.17: the method recovers the direction of the association and understates its size. The simulation assumes no censoring after a simulated buzz, because when an AI teammate would have answered is unobserved.

Check (dense outcome)	Hidden-target estimate
Primary	1.17 [0.56, 1.79]
Refit on 200 question resamples	1.18 [0.53, 2.06]
One human ability per game room	1.16 [0.53, 1.80]
Original grades (no alias repair)	1.06 [0.40, 1.72]
Difference prior s.d. 1.0 (primary 1.5)	0.69 [0.32, 1.07]
Difference prior s.d. 2.5	1.87 [0.92, 2.81]
Slot smoothing halved	1.15 [0.55, 1.74]
Slot smoothing doubled	1.14 [0.53, 1.75]
Simulation recovery (true 1.17)	0.60 [0.34, 0.90]

Table 6: The hidden-target coefficient stays positive, with its interval above zero, whenever the difficulty model is refitted; simulations recover its sign but understate its size. Intervals are 95% question-clustered intervals, except for the resampling row (median and percentile interval over 200 refits) and the simulation row (mean and percentile interval over 50 simulated data sets). Uncertainty in the difficulty fit does not explain the association away.

A.8 Held-Out Prediction

Adding the characteristics does not improve held-out prediction of adversarialness. Five-fold ridge prediction (regularization fixed at $\alpha = 10$, scaling fitted only on training folds) holds out entire source

questions, then separately entire author groups. Adding all traits worsens root mean squared error from .757 to .772 for held-out tossup questions and from 1.097 to 1.129 for bonus questions. For held-out authors, the tossup error improves from .771 to .765, while the bonus error worsens from 1.142 to 1.187. The predicted scores come from one fit of the difficulty model on all questions, including the held-out ablations, so this only tests how well the characteristics predict those scores.

A.9 Recognition Without the Requested Answer

Wrong answers on bonuses reveal two further patterns connected to the task characteristics (Figure 11). The cases below suggest possible failures but do not establish effects across the collection.

Systems return the source instead of its subject.

A bonus shows an adaptation of *Washington Crossing the Delaware* and asks for the original subject (Figure 11a). Five of ten systems return the painting’s title rather than *George Washington*; humans answer unaided correctly in all five opportunities. The title answers show the systems recognize the source but do not name the requested person, which the artwork and alteration labels do not capture.

Systems fail to complete a transformation. In the Peru bonus (Figure 11b), none of the ten systems gives the accepted answer, whereas humans answer correctly in two of four opportunities. Recognizing the pictured word is only the first step of this *symbolic manipulation*, and wrong answers alone do not show which step each system missed. Similarly, half the systems answer the kea rebus with *Kiwi*, the country’s best-known bird, instead of decoding *key* + *A* (Figure 11c). Symbolic tasks do not uniformly favor humans, however: six of ten systems decode a composite rebus for Boston that no human team answers unaided (Figure 11d), consistent with the null task-demand associations in Figure 5.

B GRACE Calibration Sensitivity

Hidden targets remain the only tossup characteristic with $q < .05$ when an adaptation of GRACE’s calibration score (CalScore; Sung et al., 2025a) replaces centered DIF as the outcome. CalScore compares model confidence and response timing with human buzzes; unlike DIF, it does not estimate ability-adjusted item difficulty. The comparison keeps the primary analysis’s audited machine

grades, compatible question versions, 560 slots from 110 tossups, and 330 parts from 110 bonuses.

Score construction. For system m , question i , and checkpoint t , let $y_{mit} \in \{0, 1\}$ be audited correctness and $\kappa_{mit} \in [0, 1]$ the recorded confidence. We set $g_{mit} = 2y_{mit} - 1$. Let h_{it} be the fraction of all compatible human episodes for i whose first human buzz is correct and occurs by t . These are aggregate first-human-buzz episodes, excluding AI teammates’ buzzes (Appendix A.1). Censored episodes remain in the denominator, but their unobserved answers are not labeled incorrect; h_{it} measures observed correct-buzz incidence under the tournament stopping protocol.

Following the published equation and its normalization, define

$$r(x) = \frac{\sigma(x) - \sigma(-1)}{\sigma(1) - \sigma(-1)}, \quad (5)$$

$$C_{mi} = 1 - r\left(\frac{1}{T_{mi}} \sum_{t=1}^{T_{mi}} (1 - h_{it}) g_{mit} \kappa_{mit}\right), \quad (6)$$

where σ is the logistic function and T_{mi} counts available unique checkpoints. We normalize separately for each system, then average systems equally within each question. For slot-level analysis, we use the single-checkpoint form of Equation 6 at each exact slot endpoint, then average systems; we do not average confidence across systems before normalization. Dense guesses can contribute after a system’s first buzz. Missing endpoints are not filled with stale guesses, and nine duplicated terminal positions created by clipping overshooting checkpoints use the last observation. Confidence is used directly, without exponentiation. Bonus parts use the same single-step formula with h equal to the fraction of opportunities answered correctly without AI consultation. This bonus construction is an extension of the incremental GRACE task.

Implementation sensitivity. We note that our released code for GRACE scoring uses the expression $1 - \mathbb{E}_t[(1 - h_{it})y_{mit}\kappa_{mit}]$, without signed correctness or the normalized sigmoid, and divides cumulative correct human buzzes by the number of eventual human buzzes. We compute this alternative separately, retaining our first-human-buzz cohort, whereas the primary adaptation above uses all compatible episodes as its denominator. Neither variant reproduces GRACE’s original benchmark scores.

Recognition is not enough

(a) Return the requested subject after recognition

Bonus packet 4, part 2



Task "Name the original subject of the painting that each of these artworks is based on."

Answer George Washington

Humans 5/5
AI 0/10

✗ Washington Crossing the Delaware 5 of 10

Demand Move from recognizing the source painting to naming its subject. A title response supports recognition but does not explicitly give the requested person.

(b) Transform a pictured word into a capital

Bonus packet 4, part 1



Task The lead-in asks for the country whose capital is an anagram of the pictured word.

Answer Peru

Humans 2/4
AI 0/10

✗ Mali, Laos, France among others

Demand *mail* → *Lima* → *Peru*. The outputs establish failure to reach the answer, but do not identify the failed step. This is coded as symbolic manipulation.

Composite rebus

(c) Decode the rebus, not the familiar animal

Bonus packet 4, part 3



Task "Based on the rebus and country shown, what is the animal?"

Answer kea

Humans 3/5
AI 0/10

✗ Kiwi 5 of 10

Demand Decode *key* + *A* and use the outline of New Zealand to name the kea. Half the systems instead name the country's best-known bird.

(d) Counterexample: a symbolic task favors machines

Bonus packet 4, part 3



Task "Name the city containing a sports stadium represented by this rebus."

Answer Boston

Humans 0/5
AI 6/10

✓ Boston 6 of 10

Demand Combine *fen* and *weigh* to recover Fenway, then name its city. Like the kea example, this is a composite rebus requiring a further relation, yet its operational gap favors machines.

Figure 11: Systems recognize what these bonus images show but stop short of the requested answer (a–c), although a symbolic task can also favor machines (d). (a) Five of ten systems name the source painting instead of its subject. (b) No system completes *mail* → *Lima* → *Peru*. (c) Half the systems answer *Kiwi* instead of decoding the rebus. (d) Six of ten systems, but no human team, decode *fen* + *weigh* to name Boston. Images are the delivered focal images, without the separate lead-in image; quoted text is verbatim, except the paraphrased task in (b). Humans: unaided correct answers over opportunities; AI: systems correct. Each bubble shows an incorrect response (a correct one in (d)) and how many systems gave it; (b) lists three of the ten responses, and the count in (c) ignores capitalization. Recognizing the image is only the first step of these tasks.

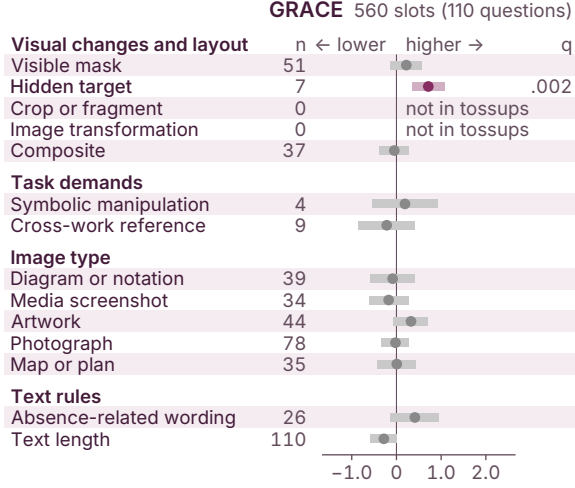


Figure 12: Hidden targets are again the only tossup characteristic that passes the multiple-testing check ($q = .002$) when GRACE’s calibration score, with all compatible human episodes in the denominator, replaces DIF as the outcome. Coefficients and 95% question-clustered intervals are in outcome standard-deviation units; colored marks indicate $q < .05$ (others gray), q is printed when below .10, and n counts questions with the characteristic. The association is not specific to IRT-based adversarialness.

Association results. We retain the primary analysis’s position-weighted cumulative tossup features, packet and position controls, separate regressions for each characteristic, question-cluster confidence intervals, and Benjamini–Hochberg adjustment within each outcome (12 estimable tossup and 14 bonus characteristics). Hidden targets remain the only characteristic with $q < .05$ under either GRACE variant for tossups; no bonus characteristic meets that cutoff (Table 7). The intervals condition on the estimated scores and do not propagate uncertainty in the human curves or recorded confidence. Figure 12 shows the GRACE coefficients, dividing each coefficient and interval by the outcome’s standard deviation; the primary DIF coefficients are in Figure 5. This rescaling aids comparison without making DIF and CalScore the same construct.

Agreement and interpretation. The primary CalScore adaptation correlates with DIF at Spearman $\rho = .785$ across tossup slots, .883 across whole tossups, and .653 across bonus parts; the released-code adaptation gives .367, .533, and .822, respectively. Whole-question DIF averages slot estimates, while whole-question CalScore averages checkpoint contributions within each system before normalization. Only 17 of 56 highest-ranked

Type	Outcome	β [95% CI]	q
Tossup	DIF	1.174 [0.557, 1.790]	.0031
	CalScore	0.168 [0.081, 0.255]	.0025
	Code variant	0.162 [0.058, 0.266]	.0309
Bonus	DIF	0.316 [−0.460, 1.092]	.5899
	CalScore	0.054 [−0.092, 0.199]	.6510
	Code variant	0.049 [−0.077, 0.176]	.8816

Table 7: Hidden targets are positively associated with all three tossup outcomes ($q < .05$) and with none of the bonus outcomes. DIF coefficients are in logits; CalScore coefficients are in score units and are not directly comparable in magnitude. Both the tossup association and the bonus null hold across outcomes.

tossup slots overlap between primary CalScore and DIF. As h approaches one, signed CalScore approaches .5 regardless of correctness: errors after humans answer receive less weight. Consequently, preservation of the hidden-target association does not imply interchangeable item rankings or constructs. We retain DIF for ability-adjusted machine-relative difficulty and use GRACE as a complementary confidence-and-timing sensitivity analysis.

Reproducibility. The anonymous repository (Section 3) provides the code, frozen intermediate observations, all scores, and complete coefficient tables. The replay reproduces the original scores and coefficients to numerical precision without new model calls.

C Authoring Analysis: Methods, Codebooks, and Examples

C.1 Reconstructing the Image-Revision Corpus

We link Firebase authoring events and saved image versions to VLM run records using session, slot, run identifiers, and timestamps. We combine pairs identified from the interface’s single-test events with pairs reconstructed from the fuller run history, deduplicating by before/after run identifiers. This yields 533 candidate image-change pairs. Of the 91 exclusions, 68 changed the stored gold answer, 9 fell outside the selected author cohort, 6 introduced an image into a blank slot rather than revising existing visual evidence, and 8 had other reasons (model/prompt scope, unresolved run links, blank after-image, case-specific review), leaving 442 pairs from 114 sessions owned by 17 authors: 246 tossup pairs and 196 bonus pairs. There are 235 distinct session–slot combinations. Repeated

tests and revisions mean that these counts do not represent independent questions.

All selected tests use the recorded model label Qwen3-VL. The retained prompt versions are `general_v1` (245 pairs), `bonus_v1` (196), and one `buzz_v1` pair (wrong-to-wrong). We therefore do not describe every retained test as a forced-answer probe. Gold-answer comparison normalizes capitalization, punctuation, and spacing, but does not judge semantic equivalence. Conversely, an unchanged answer key is not evidence that an image still supports that answer.

We retain pairs with intervening runs, recorded editorial work, image uploads during inference, or possible slot transfers. Such warnings are available at the record level rather than treated as automatic exclusions. Thus, a pair describes an observed revision interval, which need not be consecutive in the complete history, and the displayed image cannot always be verified as the exact image consumed by inference. The session owner also need not be the person who edited the pixels. Stored final question text supplies context where historical wording is unavailable; it must not be read as a verbatim prompt at both endpoints.

C.2 Developing and Applying the Codebooks

Initial model-assisted descriptions compared before/after images with available question context and stored answers, using Meta Muse Spark 1.3 Contributor⁴ and GPT-6 Astra. The descriptions supported an initially finer set of groups, which we refined interactively through inspection and discussion into five observable revision categories (Table 8). The saved visual review covers all 438 retained distinct image pairs on 56 contact sheets with targeted larger inspections. Four visual pairs each occur in two distinct run pairs; shared visual codes are expanded to all 442 test pairs, while retaining their individual responses and grades.

The five clue-level strategies (Table 9) were developed from close readings of recorded correct-to-wrong cases and then applied to every selected pair. For each pair we record the observed change, proposed surviving route, clue-level labels, interpretation support, and edit direction. The all-pair pass uses the saved visual observations and response/context records, incorporates the detailed correct-to-wrong audit, and directly reinspects 24 selected image pairs to resolve boundaries. These

are assistant-produced annotations with researcher-guided codebook refinement, not independent human coding. We have neither inter-rater reliability measurements nor a held-out human annotation set; the model-assisted procedure has not been validated against such a set.

Clue-level labels describe only what the *revision changes*. Resizing an unchanged relational diagram does not automatically add a relational-strategy label. A pair can, however, require both composing a clue within the image and retrieving an association beyond it (Figure 14c). Restoring a label can reverse cue withholding; we retain that direction rather than calling the restoration an attack.

At the record level, 202 interpretations are supported and 240 are conditional. “Supported” concerns the evidence for the interpretation, not measured human answerability; “conditional” marks uncertain source identity, clue sufficiency, or alignment with historical wording. The categories were developed with access to failures, and the review was not blinded to outcomes. In particular, the tempting-alternative label can be easier to recognize once the model has produced a distractor. This makes its high recorded failure rate unsuitable as an unbiased estimate of strategy effectiveness.

Figure 13 shows how one observable revision can reshape the route to the answer in several ways, making both codebooks necessary. Masking mostly withholds an easy cue (104 of 106 pairs) but often also creates relations within the image (44), switching the identifying clue mostly introduces an intermediate clue (96 of 130), and restyling almost always preserves content (127 of 131).

C.3 Recorded Outcomes and Uncertainty

Figure 6 conditions on initially correct tests, which are $80/(147 + 80) = 35.2\%$ overall. Initially wrong tests cannot contribute to this denominator. Table 10 preserves all four outcomes (CC, CW, WC, WW) so that failure-to-recovery changes are also visible.

For each plotted row, we draw 5,000 bootstrap samples of sessions with at least one initially correct eligible pair, preserving all eligible pairs within each sampled session. We recompute the rate and report its 2.5th and 97.5th percentiles. Resampling is performed independently by row. The intervals describe variation across sessions. They do not correct outcome-aware coding, repeated authors, uncertain image timing, or grading errors, nor do overlapping marginal intervals constitute a formal

⁴<https://openrouter.ai/meta/muse-spark-1.3-contributor>

Observable revision	Coding boundary	Example
Mask or restrict evidence	Cover, crop, or remove an identifying cue, object, region, or guidance. The operation concerns what becomes unavailable; it does not establish that the remaining task is solvable.	Label masking (Figure 7a)
Re-express the same clue	Substitute another view, example, or representation of a retained clue. This differs from switching to a different entity or association as the identifying clue.	Wider stage view (Figure 7c)
Switch the identifying clue	Replace the evidence used for identification with a different clue to the stored target. A new image source alone does not determine this label.	Blood cells to sickle (Figure 7b)
Resize, rotate, or restyle	Change scale, position, orientation, color, or rendering of existing content. This observable category does not by itself guarantee semantic equivalence.	Changes to scale or orientation
Reveal or supplement evidence	Restore obscured material, enlarge the available scene, or add labels, pointers, or other guidance. These edits need not be attempts to increase difficulty.	Added labeling

Table 8: Primary revision categories, in reporting order (counts in Table 10). Each of the 442 pairs contributes once. Mixed edits retain secondary observations, but these do not duplicate the primary counts.

Clue-level strategy	Proposed change in the answer route	Example
Withhold an easy cue	Remove an immediate recognition route while retaining a specific alternative. Record what survives and how it supports the target; masking alone is insufficient.	Focus/periphery after masking (Figure 7a)
Tempting alternative	Leave a plausible alternative answer in tension with more diagnostic evidence. Name both the alternative and the evidence distinguishing it; a wrong answer alone is insufficient.	Docs versus Sheets controls (Figure 14b)
Intermediate clue	Require a recoverable knowledge link from an interpreted clue to another entity or concept: “This is X ; what requested target is associated with X ?”	Fruity breath to diabetic complication (Figure 14c)
Relations within the image	Require interpreting roles, connections, correspondences, or combinations among depicted elements: “What does their arrangement establish?” The crucial relation is encoded within the depiction.	Fruit + breath as a symptom clue (Figure 14c)
Content-preserving presentation	Change how identifying information is presented while retaining the content needed for the same answer. Verify content preservation separately from the observable operation.	Stage view (Figure 7c; conditional)

Table 9: Clue-level codebook (counts in Table 10). A pair can receive several labels; 425 pairs receive at least one, and 17 reviewed pairs remain unresolved. Labels include conditional interpretations and reversals that restore a route. The short names match Figure 6b, Table 10, and Figure 13.

test of category differences.

Author concentration further limits interpretation. One session owner contributes 106 of the 125 relation-labeled pairs, so the observed frequency and outcomes of relational strategies need not generalize to another author or subject distribution.

C.4 Close Reading of Correct-to-Wrong Cases

For each of the 80 selected CW records, the qualitative audit records surviving evidence, a possible solution route for a knowledgeable human, and the model’s observed answer/explanation. Twenty-seven records have concrete surviving routes in that audit, 44 have conditional routes, and nine do not establish one. These judgments are distinct from support for a clue-level label and not human response measurements. The three recurring patterns

in Section 5.2 overlap and do not cover all records: familiar guesses overriding diagnostic evidence occur in 42, missed clue-to-answer connections in 43, and recognition changes across presentations in 15; 12 receive none of these patterns.

C.4.1 Scope of the Main-Paper Examples

Attention (Figure 7a). The explanatory labels are obscured while a sharply focused region and blurred periphery remain. Together with the available psychology-to-machine-learning context, this is a plausible analogy for selective attention. The recorded answer changes from “Attention Mechanism” to “Neural networks.” The latter explanation relies on the broad clue of psychological inspiration rather than accounting for the specific arrangement. The image remains an analogy for attention.

Observable revisions → clue-level strategies 442 test pairs, 578 label links

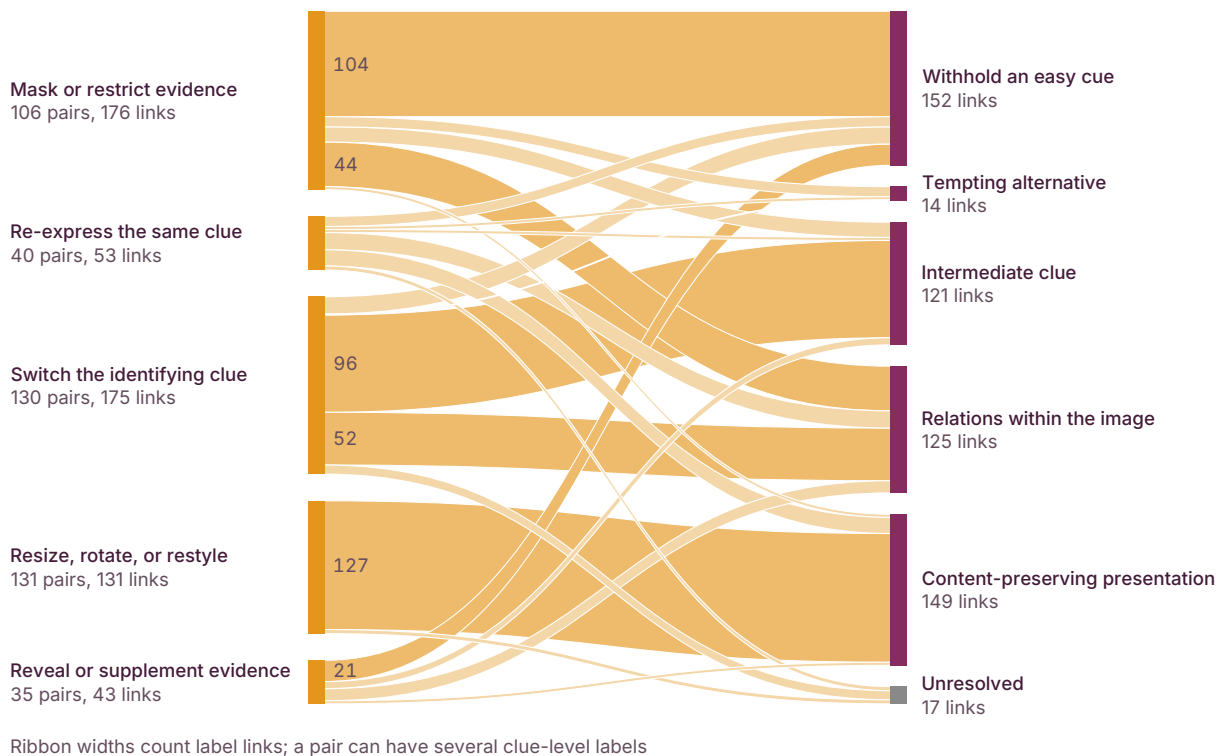


Figure 13: Most observable revisions map to several clue-level strategies; only restyling maps almost entirely to one (content-preserving presentation). Each ribbon counts pairs with that category–strategy combination; a pair with several labels contributes several links, so 442 pairs give 561 clue-level assignments plus 17 unresolved links (578 in total). Block widths count links, and left labels also report distinct pairs; conditional and restoration interpretations are included. What an edit changes in the image does not determine how it changes the route to the answer.

Hemoglobin (Figure 7b). The image changes from blood cells to a sickle, and the recorded response changes from “hemoglobin” to “DNA.” The after-explanation itself mentions sickle-cell disease and variant hemoglobin, then chooses DNA. The object supplies an indirect association, but does not uniquely identify the requested molecular level. The explanation also describes textual clues that differ from the saved final-session clue. We therefore treat preservation of the hemoglobin route as conditional on the historical question, rather than declaring the after-image independently sufficient.

Aida (Figure 7c). The images change from a closer Egyptian-themed theatrical view to a wider stage view. The available final-session clue asks for an opera known for overwhelming spectacle. Before revision the model answers “Aida”; afterward it describes a theatrical reenactment of an ancient Egyptian scene, mentioning period costumes, statues, and hieroglyphic reliefs without naming an opera. A knowledgeable reader could connect the setting and question to *Aida*, but nei-

ther the source identity nor unique answerability follows from the image alone. The clue-level presentation label is therefore conditional: it concerns recognition across depictions with surviving clues, rather than a verified content-identical transformation. This pair also spans intervening runs, so it is an illustrative observed change, not a controlled viewpoint intervention.

C.4.2 Diagnostic Evidence and Competing Answers

Italy versus Austria (Figure 14a). The revision covers a label with “Skiarea Val Comelico” on a ski map, but the name remains in its upper legend; the map is also slightly repositioned/rescaled. The recorded answer changes from “Italy” to “Austria.” The after-explanation places Val Comelico in Italy but asserts that the main resort is Austrian based on other place names. A geographically knowledgeable reader could use the surviving specific locations to identify Italy instead of focusing solely on the German names. The response thus recognizes a location cue but dismisses it.

Category or strategy	All	CC	CW	WC	WW	CW rate	95% interval
<i>Observable revision (one primary label)</i>							
Mask or restrict evidence	106	49	26	4	27	34.7%	23.4–47.7
Re-express the same clue	40	17	11	2	10	39.3%	22.2–57.7
Switch the identifying clue	130	36	28	24	42	43.8%	32.7–58.0
Resize, rotate, or restyle	131	36	12	21	62	25.0%	13.9–37.2
Reveal or supplement evidence	35	9	3	8	15	25.0%	0.0–50.0
<i>Clue-level strategy (all labels; rows overlap)</i>							
Withhold an easy cue	152	64	34	15	39	34.7%	24.0–47.2
Tempting alternative	14	3	11	0	0	78.6%	50.0–100.0
Intermediate clue	121	39	23	20	39	37.1%	27.3–49.0
Relations within the image	125	39	30	10	46	43.5%	30.0–60.6
Content-preserving presentation	149	42	21	22	64	33.3%	23.8–43.4
Unresolved	17	5	5	3	4	50.0%	—

Table 10: Correct-to-wrong rates range from 25.0% to 43.8% across categories and strategies, except tempting alternatives (78.6%), and most rows also include wrong-to-correct changes. CC, CW, WC, and WW count pairs by the model’s correctness before and after the revision; intervals are 95% session-bootstrap percentages. Unresolved records are reviewed cases without a demonstrated strategy, included for accounting. Revisions also fix the model’s answers, which rates conditioned on initially correct tests do not show.

Google Sheets versus Google Docs (Figure 14b).

Masking some menu controls changes “Google Sheets” to “Google Docs,” even though currency, percentage, decimal, and number-format controls remain. The after-explanation acknowledges these remaining controls’ association with spreadsheets but calls the font dropdown, a control common to both Sheets and Docs, the most telling clue. A plausible human route is to distinguish app-specific controls from controls shared between applications. The response documents this conflict more directly than a wrong label alone could.

C.4.3 Multi-Element Composition Clues

Fruit and breath (Figure 14c). The after-image juxtaposes fruit with visible exhaled breath; the available question concerns diabetic complications. Our interpretation composes these into “fruity breath,” then links the symptom to diabetic ketoacidosis, an association documented by [Centers for Disease Control and Prevention \(2024\)](#). The model instead interprets fruit as healthy eating and the other scene as a lifestyle/environment cue, answering “Diabetic retinopathy.” Composing the panels is strategy 4 (relations within the image), and retrieving the symptom–condition association is strategy 3 (intermediate clue).

D Further Interface Features

Authors wrote bonuses in a separate editor. A bonus starts with a lead-in shared by all parts; each part has its own image, clue and answer, plus an optional reference image that players see; it is not

needed for the answer, and neither the authoring-time model nor the shared-task systems receive it. Authors could test one part, or all parts in one run, with a model of their choice; for each part, the model sees the lead-in text and that part’s clue and image and must answer, as in the single-slot test. The editor also let the authors generate a model-attention heatmap overlay for an individual slot from its current image and clue. A separate research viewer, built after data collection, places each slot’s uploaded image versions, clue saves and logged model runs on one timeline and compares the author’s submitted clue text with the final text; it does not re-run models. Of the 276 submitted questions, 97.1% were tested at least once: 52.5% with a single-slot test, 87.0% with a buzz test (for bonuses, a run of all parts), and at least 8.3% with a text-only test (only each slot’s latest text-only result was saved).

E Additional Related Work

Modality shortcuts. NaturalBench pairs each question with two images that call for different answers, so a model that ignores the image cannot succeed; across 53 VLMS, over 90% of them trail human accuracy by 50–70% ([Li et al., 2024](#)). M3IRT decomposes model ability and item difficulty into image-only, text-only, and cross-modal components and flags questions answerable from one modality ([Uebayashi et al., 2026](#)). Although our interface has a text-only test, we do not evaluate the final questions with one modality removed (see Limitations); a cross-modal decomposition of

slot-wise IRT would separate visual from textual demands.

Adaptive evaluation. IRT also supports choosing informative test items adaptively ([Hofmann et al., 2025](#)). Slot-wise estimates add a human reference to such item parameters, so adaptive evaluation could prefer items that separate machines from skilled humans.

(a) Mask location labels; retain one in the legend

Question (final-session context) Given a map of a ski resort, give the name of the country it is in. Named for a battlement, this nation's ski resort pictured here is in the north of this country.

Before ✓ Italy



After ✗ Austria



Recorded explanation (excerpt)
"The map shows ski resorts such as "Ski Area Val Comelico," "Kreuzbergpass," and "Fischental Val Fiscalina," which are located in the Italian Alps. ..."

Recorded explanation (excerpt)
"...The map also includes 'Ski Area Val Comelico', which is in Italy, but the main resort shown is clearly Austrian. ..."

(b) Cover controls; retain diagnostic number formats

Question (final-session context) Google's productivity suite includes editors for documents, spreadsheets, presentations, and more — each with a distinct menu structure. Given a screenshot or description of a menu bar, identify which Google app it belongs to.

Before ✓ Google Sheets



After ✗ Google Docs



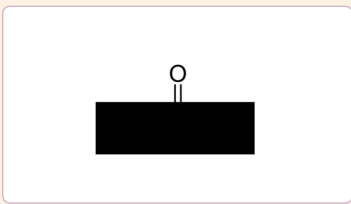
Recorded explanation (excerpt)
"...The presence of a 'Σ' (sigma) icon, which universally represents the SUM function and is a hallmark of spreadsheet applications. ..."

Recorded explanation (excerpt)
"...The inclusion of currency symbols (\$, %, .00) and number formatting (123) ... but more commonly associated with Google Sheets. However, the most telling clue is the presence of a 'Default...' font dropdown and a font size selector (10), which are standard in Google Docs. ..."

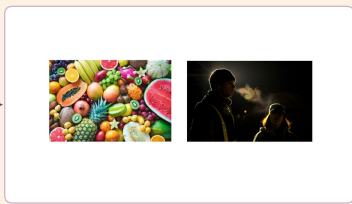
(c) Change a chemical clue to fruit + breath

Question (final-session context) Answer these questions about complications related to diabetes based on the associated images. This complication is more common for one type of diabetes than the other.

Before ✓ Diabetic ketoacidosis



After ✗ Diabetic retinopathy



Recorded explanation (excerpt)
"...acetoacetate, which is a ketone body. ... The accumulation of these ketone bodies leads to metabolic acidosis, which is the hallmark of diabetic ketoacidosis (DKA). ..."

Recorded explanation (excerpt)
"...fruits, which are often associated with healthy eating and diabetes management. ... the context of diabetes complications and the emphasis on health and environment suggest this is the most plausible answer. ..."

Figure 14: In each revision, answer-relevant evidence survives, yet the model's answer turns wrong. (a) Masking location labels leaves Val Comelico readable in the legend; the model acknowledges Italy yet answers Austria. (b) The edited toolbar keeps spreadsheet number-format controls, yet the model answers Google Docs from a shared font dropdown. (c) The model reads the fruit-and-breath composition as diet and environment rather than as fruity breath. Question text is the final saved session text (test-time wording was not versioned); explanation excerpts are verbatim, with ellipses marking omissions; images are saved versions, with a common blank-margin crop for (b). The model's own explanations show it weighing or connecting the evidence incorrectly.