Seeing Is Solving: MLLMs, Reasoning, and Refusal in Visual Math

Ethan Croteau
Worcester Polytechnic Institute
Worcester, MA
ecroteau@wpi.edu

Neil Heffernan
Worcester Polytechnic Institute
Worcester, MA
nth@wpi.edu

Many middle–school math problems are image-dependent: the diagram or graph carries essential information. This matters for intelligent tutoring and accessibility, where systems must reason over figures and also decline responsibly when figures are missing. We evaluate six contemporary multimodal large language models (MLLMs)—three reasoning models and three non-reasoning models—on 376 Illustrative Mathematics (IM) *items* labeled as image-role *Required* (the figure contains task-critical information not recoverable from text alone without added assumptions). Each model attempts every item three times with and without the figure under a shared prompt and scoring protocol. To reduce image-role label subjectivity, we classify items as not *Required* when they are solvable from text alone without additional assumptions.

With images, the top-performing reasoning models achieve accuracy in the mid-50%, while non-reasoning models fall in the mid-30s to low-40s. Without images, models overwhelmingly refuse rather than guess, with only rare correct-by-chance answers. Models show moderate agreement on which items are solvable, and we release two benchmark subsets of items solved consistently across models. A qualitative audit of 83 items shows that visual misreading is the dominant failure mode for non-reasoning models, while reasoning models more often produce correct answers accompanied by adequate explanations.

These results suggest tutoring systems should gate automated scoring and learner-model updates on visual-evidence availability and use scaffolds that require explicit visual-evidence binding before algebra. For accessibility, systems should treat no-image refusals as missing-context signals and elicit the figure or a structured description, enabling description-substitution experiments. We release code, prompts, and summary artifacts for replication. **Code and data:** https://osf.io/ct7bg/

Keywords: educational data mining, multimodal large language models, image-dependent math, refusal, reasoning, accessibility

1. Introduction

Many middle-school mathematics problems are *image-dependent*: they rely on diagrams, graphs, or other figures to specify quantities and relations that are not recoverable from text alone. Robust support for such items is critical for both intelligent tutoring systems (ITS) and accessibility. Decades of ITS research emphasize stepwise problem solving, mastery learning, and data-driven adaptation (Corbett and Anderson, 1994; Anderson et al., 1995; Woolf, 2008; VanLehn, 2011; Pelánek, 2017), and platforms such as ASSISTments instantiate these ideas at scale

in classroom-authentic settings (Heffernan and Heffernan, 2014; Roschelle et al., 2016; Selent et al., 2016). In visual math, effective support requires binding labels and givens in the figure to symbols in the solution, checking scales and units on graphs, and making geometric relations explicit–practices aligned with diagram comprehension and multimedia learning principles (Larkin and Simon, 1987; Sweller et al., 1998; Ainsworth, 2006; Mayer, 2020). Learners who cannot see or access images depend on alternative representations such as alt text, extended descriptions, audio, or tactile graphics. Accessibility guidelines for STEM materials stress that these alternatives must preserve task-relevant structure–naming entities, encoding geometric and quantitative relations, and clarifying scales and units–while managing cognitive load (DIAGRAM Center, 2015; W3C Accessibility Guidelines Working Group, 2024; W3C Web Accessibility Initiative, 2022; Braille Authority of North America (BANA), 2022; Marriott et al., 2021; Benetech, 2019; ASSETS 2025 Organizing Committee, 2025; Trewin, 2019). When tutors misread visual information or alternative descriptions omit key visual relations, students may receive misleading guidance or be excluded from substantial portions of the curriculum.

Recent multimodal large language models (MLLMs) show strong performance on visual-math benchmarks, but these evaluations rarely capture deployment-relevant reliability signals on curriculum-authentic, image-dependent items. For tutoring and accessibility, the key questions are run stability across stochastic runs, calibrated refusal when a required figure is missing, and cross-model agreement on which items are reliably solvable.

For educational use, several evaluation gaps remain, particularly for image-dependent items. Empirical studies of MLLMs on visual math report encouraging headline accuracies (Feng et al., 2025; Sáez et al., 2025) but rarely examine *run stability* (repeatability across multiple stochastic *runs*, i.e., independent attempts on the same item under the same prompt), provide limited evidence about *calibrated refusal* (explicitly declining when required information such as a figure is missing), and seldom measure *cross-model agreement* (whether different models tend to succeed on the same items) on which items are reliably solvable. Coverage of middle-school curricula and classroom-authentic problems is uneven, and explanation quality is often assessed informally or via coarse proxies rather than targeted audits of visual evidence binding, warranting, and language calibration (Kadavath et al., 2022; Ouyang et al., 2022; Chen et al., 2025). Parallel lines of work document persistent failures in visual reasoning, including perceptual misreads and relational leaps (Rudman et al., 2025; Yin et al., 2025; Tang et al., 2025). Recent diagnostic analyses further suggest that many apparent diagram "reasoning" failures originate in upstream perception limits–especially fine-grained grounding–leading models to rely on text-driven shortcuts rather than visual evidence (Sun et al., 2026). Recent benchmarks and educational analyses emphasize the need to look beyond accuracy toward faithfulness in error handling and explanation behavior (Yan et al., 2024; Zhang and Graf, 2025; Xia et al., 2025). Yet this line of work is rarely connected to curriculum-authentic middle-school items, explicit image-dependence labels (REQUIRED/USEFUL), or agreement and stability measures that are central for tutoring and accessibility decisions.

1.1. POSITIONING AND NOVELTY

Our goal is complementary to recent work that introduces new architectures or large benchmark suites for visual math: rather than proposing a new model or dataset at scale, we contribute an evaluation framework tailored to educational deployment decisions on curriculum-authentic items. In particular, we operationalize and audit an image-role taxonomy that determines which

items are intrinsically image-Required, measure robustness under stochasticity via repeated runs and item-level aggregates, and test calibrated behavior under missing modality by pairing with-image and without-image conditions. These elements are central for tutoring and accessibility, where systems must both solve reliably *and* decline responsibly when essential visual information is absent.

This paper addresses these gaps through a systematic two-condition evaluation of six multimodal LLMs–three reasoning-focused models (grok-4, o3, o4-mini) and three non-reasoning models (gpt-4.1, gpt-4o, grok-2-vision)–on a curated set of 376 image-Required middle-school items from Illustrative Mathematics (IM) (Mathematics, 2019). We first apply a dependence-oriented taxonomy of image roles (REQUIRED, USEFUL, NOT REQUIRED, INSUFFICIENT) to 541 image-containing items, then perform a conservative, human-judged refinement focused only on REQUIRED candidates: items are downgraded REQUIRED→USEFUL only when the text alone supports a unique solution with no added assumptions. Each model then attempts every REQUIRED item in three independent *runs* per condition: *with-image* (problem text + original figure) and *without-image* (same text with the figure removed). The *without-image* ablation probes whether models appropriately abstain when required information is missing rather than guess–an important property for instructional and accessibility settings where figures may be unavailable. As a controlled ablation, this design directly targets robustness (run stability; repeatability across stochastic runs), calibrated refusal under missing modality, and cross-model agreement on which items are reliably solvable.

1.2. KEY TERMS

We use *item* for one problem, *run* for one independent attempt, *with-image/without-image* for whether the figure is provided, *majority-correct* ($\geq$2/3) and *always-correct* (3/3) for item-level robustness, *refusal* for explicit inability due to missing information, *run stability* for within-item consistency across runs, and *cross-model agreement* for overlap in which items models solve.

1. **RQ1 (Accuracy and run stability with images).** How accurate and run-stable are these models on image-Required items when figures are provided, and how do reasoning and non-reasoning models compare?

2. **RQ2 (Calibration without images).** When images are removed on image-Required items, do models appropriately refuse rather than guess, and how often do we observe guess-wrong versus lucky-correct answers?

3. **RQ3 (Cross-model agreement).** To what extent do models agree on which image-Required items are solvable with images, and what overlap exists in majority-correct and always-correct items across models, including items solved consistently by all six?

4. **RQ4 (Qualitative error patterns).** Which visual reasoning and explanation errors persist even when images are available, including cases where the final answer is correct but the reasoning is incorrect or unsupported, and how do these error patterns differ across model families?

Our contributions are threefold. First, we provide a stability- and calibration-focused evaluation protocol for curriculum-authentic image-Required items, quantifying item-level robustness (majority-/always-correct), refusal behavior, and uncertainty via bootstrap confidence intervals

and family-level permutation tests (RQ1-RQ2). Second, we analyze cross-model agreement on solvability and identify two reproducible benchmark subsets–61 *ultra-gold* items solved 3/3 by all six models and 76 *gold* items solved by all six at majority-correct–supporting future diagnostics and comparisons beyond single-shot accuracy (RQ3). Third, we conduct a focused qualitative audit using a deterministic contrastive design on the exhaustive subset of REQUIRED items where the two model families separate, enabling matched "winner vs. loser" comparisons on the same problem content and revealing distinct family-linked failure modes (RQ4). Although we do not propose a new model architecture, these outputs are designed to enable improvement: they identify which items are safe candidates for automated support versus those requiring guardrails, and they motivate concrete system interventions (e.g., scaffolds that enforce visual evidence binding and refusal-first/description-substitution strategies for accessibility).

2. RELATED WORK

2.1. MULTIMODAL LLMS FOR VISUAL MATH

Multimodal Large Language Models (MLLMs) accept images alongside text (Yin et al., 2024) and increasingly perform well on visually grounded mathematical reasoning tasks such as diagram interpretation, graph reading, and geometry (Yan et al., 2025; Yang et al., 2023). Alongside instruction-tuned and math-specialized variants that aim to better ground symbolic steps in visual evidence and reduce errors like misreading scales or diagram relations (Shi et al., 2024; Wang et al., 2025), benchmark suites curate image-centric problems and report aggregate accuracy under standardized prompts (e.g., MathVista, MathVerse) (Lu et al., 2024; Zhang et al., 2024), with diagnostic ablations such as MathFlow (Chen et al., 2025), which isolate perceptual subskills and reveal persistent difficulty when visual information is non-redundant. Recent benchmarks also probe specific failure modes: VisioMath studies "visual-option" multiple-choice items where answer choices are images that can differ only subtly, and option-shuffling tests suggest models may rely on positional heuristics rather than robust text–image alignment (Li et al., 2026); EduVisBench instead evaluates the *generation* of pedagogically effective visualizations via a multi-dimensional rubric and introduces a multi-agent method (EduVisAgent) to improve such outputs (Ji et al., 2025). Prior work reports encouraging headline accuracies on visual-math evaluations (Feng et al., 2025; Sáez et al., 2025) and, separately, studies perception diagnostics or pedagogical visualization generation; in contrast, we evaluate whether MLLMs can *use provided figures* to solve curriculum-authentic middle-school items and report deployment-relevant reliability signals (Sec. 3.1).

2.2. INTELLIGENT TUTORING SYSTEMS AND VISUAL PROBLEM SOLVING

ITSs emphasize stepwise problem solving, mastery learning, and data-driven adaptation (Corbett and Anderson, 1994; Woolf, 2008; VanLehn, 2011). Classic approaches (e.g., model tracing and knowledge tracing) represent student knowledge at the skill level and intervene with hints, feedback, and worked examples (Anderson et al., 1995; Koedinger and Corbett, 2006; Pelánek, 2017). Modern platforms such as ASSISTments apply these ideas at scale, enabling item-level analytics, randomized experiments, and rapid iteration on instructional policies (Heffernan and Heffernan, 2014; Roschelle et al., 2016; Selent et al., 2016).

Visual math introduces challenges that strain text-only (without-image) pipelines. Diagrams and graphs often encode quantities and relations that must be read before symbolic work can

proceed. From a learning-sciences perspective, effective support requires binding labels and givens in the figure to symbols in the solution, checking scales/units on graphs, and making geometric relations explicit–practices aligned with multimedia learning and cognitive load principles (Larkin and Simon, 1987; Sweller et al., 1998; Ainsworth, 2006; Mayer, 2020). For classroom deployment, tutors must also handle missing or unreadable figures responsibly, preferring clarification or refusal over speculation.

2.3. Accessibility: Alternative Representations of Visual Content

Accessible STEM materials require alternatives to diagrams and graphs that preserve task-relevant information while managing cognitive load (DIAGRAM Center, 2015; Mayer, 2020; Braille Authority of North America (BANA), 2022; W3C Accessibility Guidelines Working Group, 2024). Common modalities include concise alt text, extended descriptions, audio narration, sonification, and tactile graphics. For mathematics, descriptions must name entities (e.g., points, segments, axes), encode relations/constraints (e.g., parallel, congruent, angle equality), and clarify scales/units on graphs; shortcomings in these dimensions impair comprehension and downstream problem solving (W3C Web Accessibility Initiative, 2022).

Standards emphasize accurate object naming, relation-centric phrasing, explicit units/tick spacing, and progressive disclosure to avoid overload (ASSETS 2025 Organizing Committee, 2025; Trewin, 2019). For quantitative graphics, tabular summaries and axis metadata enhance non-visual access (Marriott et al., 2021); for geometry, unambiguous verbalization of markings and constraints is critical (Benetech, 2019). Where appropriate, structural markup (e.g., MathML for expressions; structured lists/tables for graph data) further improves machine- and human-readability.

2.4. Explanations and Pedagogy in Math Learning

For educational use, how an answer is communicated matters as much as whether it is correct. The learning sciences highlight key ingredients of effective mathematical explanations: clear step coverage, warranted inferences, and cognitive load management via signaling and structure (Mayer, 2020; Sweller et al., 1998). In tutoring contexts, explanations play diagnostic and instructional roles, helping students trace which relations justify each step, pinpoint errors, and generalize principles; this demands faithfulness to the underlying problem state (VanLehn, 2006).

Recent work on LLMs assesses the pedagogical quality of generated rationales, yielding mixed results: models produce fluent, step-by-step text but often elide justifications, introduce unwarranted assumptions, or skip intermediate checks (e.g., assuming unstated diagram symmetries without evidence) (Shi et al., 2025). Chain-of-thought "scratchpad" prompting can structure intermediates but falls short on faithfulness or correctness, necessitating independent rationale validation via verifier frameworks (Nye et al., 2021; Cobbe et al., 2021). Emerging strategies like Pedagogical Chain-of-Thought (PedCoT) – a prompting technique for error-guided reasoning–show promise in flagging symbolic slips or visual misreads (Jiang et al., 2024), but remain untested at scale for visual tasks. These issues intensify in visual math, where useful explanations must tie symbolic operations to depicted relations (e.g., citing why angles are equal based on parallel markings), make scales and units explicit on graphs, and reference specific visual evidence for each step.

Despite these advances, gaps persist in evaluating LLM explanations for image-dependent items, including explicit evidence binding to diagram elements (entities, relations, scales) (Nobre et al., 2024), complete warranting of relations (e.g., parallel lines $\Rightarrow$ alternate interior angles) (Hwang et al., 2024), load-aware structure via signaling and progressive disclosure (Mayer, 2020), and calibrated language with hedging or refusal for missing info (Kadavath et al., 2022; Ouyang et al., 2022; Chen et al., 2025). Such shortcomings highlight the need for targeted audits of persistent failure modes (e.g., missing warrants for geometric relations, scale mishandling) to inform tutoring scaffolds.

2.5. Qualitative Error Analysis in Visual Reasoning

Recent audits of MLLMs reveal taxonomies of persistent errors in visual math reasoning, such as perceptual misreads in diagrams and relational leaps, even with images provided (Rudman et al., 2025). Automated frameworks classify these into categories like data extraction and symbolic slips, showing reasoning models mitigate but do not eliminate family-specific issues in geometry and charts (Yin et al., 2025; Tang et al., 2025). Diagnostic evidence suggests many downstream "reasoning" failures on diagrams trace to upstream perception limits. For example, Sun et al. (2026) introduce MATHEMETRIC, a perception-isolating suite (1,198 images; 1,609 questions) spanning shape classification, object counting, relationship identification, and object grounding, and report near-zero accuracy on fine-grained grounding for general-purpose MLLMs. They further characterize a "blind faith in text" pattern, where weak diagram perception leads models to rely on textual shortcuts rather than visual evidence. Complementary audits and educational analyses emphasize evaluating faithfulness and error handling beyond top-line accuracy (e.g., ErrorRadar (Yan et al., 2024); (Zhang and Graf, 2025; Xia et al., 2025)). We extend these lines by examining how such perception-linked failures manifest on curriculum-authentic middle-school *image-Required* items, and by quantifying run stability, calibrated refusal under image removal, and cross-model agreement on solvability.

2.6. Image-Role Taxonomies and Reliability

Determining whether a figure is Required, Useful, or Not Required (plus an Insufficient catch-all) is partly subjective and thus benefits from explicit taxonomies and reliability checks. Prior work in educational data mining recommends concise, operational definitions, multiple independent coders, and transparent reporting of agreement (e.g., Cohen's κ, Krippendorff's α) with adjudication procedures (Cohen, 1960; Krippendorff, 2011; Eagan et al., 2020). Agreement statistics are sensitive to class prevalence and unit of analysis, so reporting chance-corrected coefficients alongside raw contingency counts aids interpretability. However, existing schemes for visual dependence in math education–such as those emphasizing epistemic roles in scientific practices (Evagorou et al., 2015) or cognitive hierarchies in visual tasks (Arneson and Offerdahl, 2018)–largely focus on representational types or instructional utility, without a structured evaluation of necessity relative to text (e.g., whether key quantitative or relational information is only available in the figure, redundantly available but helpful, or fully recoverable from text alone). This gap motivates the need for a dependence-based taxonomy tailored to visual-dependent math problem solving.

3. METHODS

3.1. OVERVIEW AND EVALUATION DIMENSIONS

We evaluate six MLLMs on IM items (Sec. 3.2) under two conditions (with-image vs. without-image) with repeated runs per item (Sec. 3.3). We report deployment-relevant reliability signals: *run stability* (consistency of correctness across runs for the same item/prompt); *calibrated refusal* (whether refusal/unsolvability behavior appropriately shifts between with-image and without-image when required information is removed); and *cross-model agreement* (cross-model concurrence on item-level solvability).

3.2. DATASET AND LABELS

We curate a dataset of middle-school mathematics items from Illustrative Mathematics (IM) (Grades 6–8, standard and accelerated). To enable automated scoring and repeated evaluation, we retain only English items with a single machine-scorable final response and exactly one pedagogical image in the problem stem. We excluded items with images only in answer options, items with multiple stem images, and items where the only "image" content is a rendered mathematical expression (e.g., MathML/LaTeX-as-image) or other non-interpretive assets that do not require visual reasoning. This yielded 541 unique image-containing, auto-scorable items, each with a stable identifier, original text prompt, and an associated image. The full list of item identifiers and image-role labels is included in the accompanying reproducibility materials.

Each item was annotated for image role using four codes: REQUIRED, USEFUL, NOT REQUIRED, and INSUFFICIENT. A single rater (first author) applied a frozen operational codebook (below), following standard guidance for reproducible coding and agreement reporting in EDM (Cohen, 1960; Krippendorff, 2011; Eagan et al., 2020). Because downstream analyses condition on these image-role labels and the Required–Useful boundary is inherently judgment-based, we treat the labels as operational categories rather than objective ground truth; accordingly, we documented decision rules in the operational codebook, froze definitions prior to the reliability audit, and report agreement and confusion matrices to make remaining uncertainty transparent.

3.2.1. Labeling workflow and consistency checks

The rater recorded labels in a structured form while viewing each item's text and stem image. We deduplicated cross-listed items to a single canonical instance per stable identifier before analysis. After an initial pass, the rater completed a second consistency pass and generated per-label review PDFs to support targeted auditing of the REQUIRED set.

3.2.2. Decision criterion at the REQUIRED–USEFUL boundary

We operationalize this boundary as necessity vs. sufficiency: REQUIRED if the image contributes at least one task-critical constraint needed for a unique solution; USEFUL if the text alone is sufficient and the image is redundant-but-helpful (see definitions below).

3.2.3. Inter-rater reliability (image-role labels)

A second rater independently coded a stratified audit set ($n = 153$) using the frozen definitions, blind to the first author's labels, model outputs, and refinement flags. Across the four labels,

agreement was 71.9% (110/153) with Cohen's $\kappa = 0.591$ (95% CI [0.492, 0.687]). Collapsing to a binary boundary (REQUIRED vs. non-REQUIRED) yielded 88.2% agreement (135/153) with $\kappa = 0.764$ (95% CI [0.657, 0.865]). Disagreements concentrated near the REQUIRED–USEFUL boundary; confusion matrices appear in Appendix D and audit-set composition in Appendix C.

3.2.4. Operational definitions (image role)

Labels reflect whether the text alone (or text+figure) is sufficient for a *unique* solution without additional assumptions.

- REQUIRED: At least one task-critical quantity, relation, label, or structural constraint is present only in the image and cannot be recovered from the text alone without inventing information.
- USEFUL: The text alone is sufficient for a unique solution, and the image provides redundant but task-relevant support (e.g., clarifying correspondences, disambiguating interpretation, organizing information, reducing cognitive load).
- NOT REQUIRED: The image is non-essential and does not materially aid solving (adds no constraints and does not clarify or disambiguate the text).
- INSUFFICIENT: Even with the image, the item is underspecified for a unique correct answer (e.g., missing units/scale/dimensions).

3.2.5. Worked exemplars

To make distinctions among REQUIRED, USEFUL, NOT REQUIRED, and INSUFFICIENT concrete, Appendix A provides one representative item for each label with a brief justification aligned to the operational definitions.

3.2.6. Refinement protocol (flag-then-review; not model-defined labels)

We performed a conservative flag-then-review pass on items initially labeled REQUIRED. Model behavior served only to prioritize human review (e.g., any without-image correct run triggered a flag); reclassification decisions were based solely on applying the operational definitions to the item text (independent of model outputs). An item was downgraded REQUIRED→USEFUL only when the text explicitly contained all task-critical quantities/relations for a unique solution without introducing assumptions. Six items were downgraded (Table 14), yielding $N = 376$ REQUIRED items (Table 1); flagged items failing the text-sufficiency criterion were retained as REQUIRED (Table 15). We release both initial and refined labels; image-dependence analyses use the final REQUIRED set.

3.2.7. Solvable without assumptions (without-image)

For the refinement pass, we define an item as solvable from text alone *without assumptions* if: all quantities, labels, and relations needed for a unique solution are explicitly stated in the text; no step requires guessing a value, selecting among multiple plausible diagram interpretations, or inferring an unstated scale/unit; and the solution does not rely on generic defaults (e.g., assuming symmetry, equal partitioning, or unlabeled correspondences). When without-image solvability depended on any such inference, the item remained labeled REQUIRED.

Table 1: Final problem classification after refinement. Percentages are of the 541 image-containing items.

Classification	Count	Percent
Required	376	69.5%
Useful	122	22.6%
Not Required	26	4.8%
Insufficient	17	3.1%
Total	541	100.0%

Table 1 reflects naturally occurring prevalence within the image-containing subset (541 items), where many figures encode non-redundant quantities or relations and thus fall under REQUIRED. Because we do not train a supervised model to predict these labels, we report prevalence and downstream evaluation results by stratum (e.g., restricting core analyses to REQUIRED) rather than applying class-balancing methods such as the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002). Accordingly, classification metrics for imbalanced label prediction (e.g., AUC/F1) are not applicable; our reported accuracy measures item-solving correctness on the evaluated item set.

3.3. MODELS AND SETTINGS

We evaluate six contemporary multimodal LLMs (Table 2). For readability, we use the short codes in all subsequent tables and figures, and we aggregate by Type (Reasoning vs. Non-reasoning) when comparing families.

Table 2: Model key (codes used throughout). Full version strings specify the exact checkpoints evaluated.

Code	Full model/version
Non-reasoning	
gpt-4.1	`gpt-4.1-2025-04-14`
gpt-4o	`gpt-4o-2024-08-06`
grok-2	`grok-2-vision-1212`
Reasoning	
grok-4	`grok-4-0709`
o3	`o3-2025-04-16`
o4-mini	`o4-mini-2025-04-16`

Each model received three independent runs per REQUIRED item under two conditions (with image vs. without-image), enabling estimation of run stability and calibrated refusal as defined in Sec. 3.1. Provider defaults governed stochasticity; repeated runs estimate stability rather than enforcing determinism. We log model/version, condition, run index, final answer, correctness, and refusal. Evaluations used paid API access with anonymized prompts; no user data was collected.

3.4. Prompting and Scoring

3.4.1. Stimuli

In the with-image condition, models received the original problem text and the associated figure. In the without-image condition, the same text was presented with the figure omitted. All prompts used a consistent instruction header and a lightweight structured response to ease parsing.

3.4.2. Correctness

A run is scored *correct* when the model's extracted final answer matches the item's reference answer under our normalization and equivalence procedure (Appendix E.4). Briefly, we apply lightweight string normalization and then test algebraic/numeric equivalence using SymPy simplification when expressions can be parsed. Reference answers in this filtered dataset are unitless, so we do not perform unit-equivalence checks or unit conversions. We do not apply numeric tolerances. Correctness scoring is fully automated and does not depend on qualitative coding.

3.4.3. Refusal

We treat a run as a *refusal* when the model sets `isSolvable=false` in the structured response. Because `isSolvable` is present for every run, we do not attempt to infer refusal from free-text language (e.g., statements like "I cannot see the image"). Refusals are reported as a separate rate but are treated as incorrect for accuracy and item-level solvability calculations (e.g., majority-correct, $\geq$2/3 runs). A representative no-image refusal example is shown in Appendix B.

3.4.4. Qualitative audit (RQ4)

We conducted a focused qualitative audit on 83 Required, with-image problems drawn from the same final Required set used in the quantitative analyses ($N = 376$). The audit set is *exhaustive* rather than randomly sampled: it contains all items on which the two model families separate under our run-level rule (at least one model in one family achieves $\geq 1/3$ correct runs, while all models in the other family achieve $0/3$). For each audited item, we compared two responses with the problem content held fixed: one *winner* response from the succeeding family and one *loser* response from the failing family. This yields a main stratum where the reasoning family succeeds and the non-reasoning family fails ($n = 68$), and a counter-stratum where the non-reasoning family succeeds and the reasoning family fails ($n = 15$).

The coding unit was a single model response (steps plus final answer). Winner responses were selected deterministically from the succeeding family by restricting to models that solved the item and breaking ties using a stable SHA-256 hash of the problem identifier; loser responses were selected by cycling deterministically across models in the failing family using the same hash-derived index. We used the first correct attempt for winners (preferring attempt 1; otherwise the first correct among attempts 2–3) and, for losers, the first attempt with parsed reasoning steps (else attempt 1). To reduce expectation effects, model names and family labels were masked in the coder view (responses randomized and stripped of identifiers; problem IDs anonymized), while the original stem and figure were shown verbatim.

Two raters performed the audit. The first author served as a domain-expert coder; a second researcher provided independent ratings on a subset. Both raters calibrated on ten practice responses using the draft codebook, discussed discrepancies, and clarified decision rules before formal coding.

The taxonomy began as seven fine-grained codes (Appendix L). Pilot double-coding indicated disagreements concentrated among fine-grained visual subtypes and near the *No-error* versus minor explanation/format boundary. For analysis, we merged visual perception failures into *Visual misread* and grouped non-visual communication failures under EXPLANATION/FORMAT, retaining REFUSED for explicit non-solves due to missing information (`isSolvable=false`; Table 7). Responses with correct answers and adequate explanations were labeled *No-error* and were not counted as errors. Each response received one primary code corresponding to the earliest failure in the reasoning chain; coders recorded the triggering span and, when applicable, the relevant figure evidence.

To concentrate reliability effort where variance was highest, we double-coded the rarer counter-stratum ($n = 15$; $15/83 = 18.1\%$) and single-coded the main stratum ($n = 68$) by the first author. On the double-coded set (merged taxonomy; nominal), inter-rater reliability for non-reasoning responses was Krippendorff's $\alpha = 0.584$. For reasoning responses in the same stratum, refusal (via `isSolvable`) matched perfectly; after merging, the remaining non-refusal reasoning codes collapsed to a single category, so α was not informative. Disagreements were adjudicated to a single primary label for reporting. Because the audit set is purposefully constructed, proportions are treated as descriptive of this contrastive set.

3.5. METRICS

3.5.1. Rationale for κ (RQ3)

For cross-model agreement on REQUIRED, with images, each model provides a binary label per item: solved (correct under majority-correct, $\geq 2/3$ runs) vs. not solved (wrong or refusal). Using majority-correct labels reduces sensitivity to within-item run noise by collapsing three stochastic attempts into a single item-level outcome.

We compute pairwise Cohen's κ on these labels because it corrects observed agreement p_o by the chance agreement p_e implied by model-specific base rates: $\kappa = \frac{p_o - p_e}{1 - p_e}$. This yields a chance-corrected measure of whether two models succeed on the same items, not merely whether they have similar accuracies.

We report solved-by-k coverage in Appendix G and the underlying contingency counts (enabling κ reconstruction) in Appendix H for transparency. To further assess set-level robustness independent of base rates, we report the Jaccard index on "solved" sets (majority-correct) for each model pair: $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, provided in Appendix I.

We predefine the following metrics:

- **Accuracy (majority-correct):** fraction of items for which the model is correct on at least 2 of 3 runs ($\geq 2/3$). Items are the unit (i.e., an item-level "solved" rate under the majority rule).
- **Consistency:** percentage of items for which correctness is identical across all three runs (all correct or all wrong), per model and condition.
- **Without-image behavior:** for model m, let N be the number of evaluated runs on Required items and let $n_{\text{ref}}^{(m)}, n_{\text{gw}}^{(m)}, n_{\text{lc}}^{(m)}$ be the counts of refusal, guess-wrong (answered and incorrect),

and lucky-correct (answered and correct) responses, respectively. We report rates

$$\mathrm{ref}_m = \frac{n_{\mathrm{ref}}^{(m)}}{N}, \qquad \mathrm{gw}_m = \frac{n_{\mathrm{gw}}^{(m)}}{N}, \qquad \mathrm{lc}_m = \frac{n_{\mathrm{lc}}^{(m)}}{N}.$$

By construction, $\mathrm{ref}_m + \mathrm{gw}_m + \mathrm{lc}_m = 1$.

- **Cross-model agreement:** pairwise Cohen's κ on item-level correctness and the proportion of items answered correctly by all models.

All core analyses restrict to the final Required set unless explicitly stated.

3.5.2. Uncertainty and family contrasts

Using items as the unit, we compute 95% bootstrap confidence intervals (10k resamples) for majority-correct solved rate, refusal rate, and item-level consistency. For family-level contrasts (Reasoning vs. Non-reasoning), we use item-level permutation tests that shuffle model-to-family assignments while preserving the 3-vs-3 split; we report the observed effect (percentage-point difference) and a two-sided p-value.

4. Results

4.1. Performance and Stability on Required (with images)

We restrict all performance analyses to the Required set ($N = 376$ items), with three runs per item (1,128 runs per model). Table 3 summarizes item-level majority-correct accuracy ($\geq$ 2/3 runs), refusal rate, and correctness consistency, each with 95% bootstrap confidence intervals (10k resamples; items as unit). Table 4 provides the corresponding run counts.

Table 3: Required, with images ($N = 376$): item-level majority-correct accuracy ($\geq$ 2/3 runs), refusal rate, and correctness consistency with 95% bootstrap CIs (10k resamples; items as unit). Refusal is the mean per-item refusal rate over 3 runs. Consistency is the fraction of items that are either 3/3 correct or 0/3 correct. Codes refer to Table 2.

Model	Accuracy (95% CI)	Refusal (95% CI)	Consistency (95% CI)
Non-reasoning			
gpt-4.1	42.6% [37.8, 47.6]	15.6% [12.2, 19.2]	83.5% [79.8, 87.2]
gpt-4o	35.4% [30.6, 40.2]	13.3% [10.1, 16.7]	82.5% [78.5, 86.2]
grok-2	34.0% [29.3, 38.8]	6.7% [4.5, 9.0]	84.3% [80.6, 88.0]
Reasoning			
grok-4	40.2% [35.4, 45.2]	31.7% [27.6, 35.8]	85.4% [81.7, 88.8]
o3	52.4% [47.3, 57.5]	19.1% [15.8, 22.5]	79.0% [74.7, 83.0]
o4-mini	57.7% [52.7, 62.8]	10.7% [8.2, 13.5]	79.0% [74.7, 83.0]

Table 4: Required, with images: run-level outcome counts out of 1,128 runs per model. Codes refer to Table 2.

Model	Correct	Incorrect	Refusal
Non-reasoning			
gpt-4.1	482	470	176
gpt-4o	425	553	150
grok-2	389	664	75
Reasoning			
grok-4	458	313	357
o3	604	309	215
o4-mini	650	357	121

Takeaway All models exhibit high run stability on Required items (consistency ≈79-85%). Reasoning models achieve the highest *majority-correct item-level solved rates* (o4-mini 57.7%, o3 52.4%), while non-reasoning models range from mid-30s to low-40s. Refusal rates with images are model-dependent and vary substantially (e.g., grok-4 highest), rather than being uniformly lower for the reasoning family (Table 3).

Uncertainty We report item-level 95% bootstrap confidence intervals (10k resamples) for majority-correct accuracy, refusal rate, and item-level consistency in Table 3 and Fig. 3; full details are in Appendix K. At the family level, we report permutation tests that shuffle model-to-family assignment (3 vs. 3 models), using items as the unit for the effect statistic; these did not detect reliable differences in majority-correct solved rate (Reasoning−Non: $+12.77$ pp, $p = 0.197$), with-image refusal ($+8.63$ pp, $p = 0.298$), or item-level consistency (-2.30 pp, $p = 0.400$).

4.2. Item overlap and robustness (Required, with images)

We next examine how broadly each Required item is solved across models under two robustness notions: always-correct (3/3 runs) and majority-correct (2/3 or 3/3 runs). Rather than reporting only run-level accuracy, this view asks: for each item, how many models solve it–none, a few, or all? The full distribution of "exactly k models solved" under 3/3 and 2/3+ appears in Appendix G, Table 19. In brief, the 3/3 criterion concentrates mass toward smaller k (stricter robustness), while the 2/3+ criterion shifts mass toward larger k (more inclusive), clarifying which items are broadly accessible to current models versus idiosyncratic to specific models.

4.3. Behavior Without Images on Required

When figures are removed on Required items, we assess calibration by splitting runs into refusal versus answering (guess-wrong or lucky-correct), and we report item-level consistency (identical correctness across three runs). Table 5 summarizes run-level counts and percentages for each model (denominators: $3 \times 376 = 1{,}128$ runs/model; 376 items/model).

Across all six models, refusals dominate: 1,076–1,115 of 1,128 runs/model (95.4%–98.9%). **Answering is rare**: guess-wrong ranges from 3–34 runs (0.3%–3.0%) and lucky-correct from 10–30 runs (0.9%–2.7%). **No-image behavior is extremely stable**: item-level consistency spans 367–375 of 376 items (97.6%–99.7%).

Notable differences are small but informative: `grok-4` shows the highest refusal rate (1115/1128; 98.9%); `o4-mini` has the most lucky-correct answers (30/1128; 2.7%); `grok-2` produces the most guess-wrong answers (34/1128; 3.0%); and `gpt-4.1` reaches the highest item-level consistency (375/376; 99.7%). Overall, models are appropriately calibrated to refuse rather than speculate when critical visuals are missing; small rates of lucky-correct answers remain and are model-dependent.

Table 5: Required, without images: run-level *Count* and *Percent*, plus item-level consistency (items with identical correctness across three runs). Codes refer to Table 2.

	Refusal		**Guess-wrong**		**Lucky-correct**		**Consistency**	
Model	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
Non-reasoning								
gpt-4.1	1102	97.7%	10	0.9%	16	1.4%	375	99.7%
gpt-4o	1092	96.8%	26	2.3%	10	0.9%	373	99.2%
grok-2	1076	95.4%	34	3.0%	18	1.6%	367	97.6%
Reasoning								
grok-4	1115	98.9%	3	0.3%	10	0.9%	373	99.2%
o3	1101	97.6%	10	0.9%	17	1.5%	374	99.5%
o4-mini	1084	96.1%	14	1.2%	30	2.7%	374	99.5%

4.4. Cross-Model Agreement

Agreement across the six models shows meaningful within-family coherence and lower agreement across families; see Table 6 and Fig. 1.

Table 6: Pairwise agreement (Cohen's κ) on *Required, with images*. Cells use model codes from Table 2; values are computed on majority-correct labels ($\geq$ 2/3 runs).

Model	gpt-4.1	gpt-4o	grok-2	grok-4	o3	o4-mini
gpt-4.1	1.00	0.68	0.59	0.56	0.58	0.46
gpt-4o	0.68	1.00	0.65	0.56	0.50	0.43
grok-2	0.59	0.65	1.00	0.54	0.45	0.39
grok-4	0.56	0.56	0.54	1.00	0.53	0.51
o3	0.58	0.50	0.45	0.53	1.00	0.64
o4-mini	0.46	0.43	0.39	0.51	0.64	1.00

Summary Agreement on which Required items are solvable with images was moderate overall (mean off-diagonal Cohen's $\kappa = 0.54$). Within-family agreement exceeded cross-family: the non-reasoning trio averaged $\kappa = 0.64$, the reasoning trio $\kappa = 0.56$, while cross-family pairs averaged $\kappa = 0.50$. The most aligned pair was gpt-4.1 vs. gpt-4o ($\kappa = 0.68$), and the least aligned was grok-2 vs. o4-mini ($\kappa = 0.39$). We also identify 76 Required items solved by *all six* models at majority-correct and 61 at always-correct (3/3), which we release as "gold" and

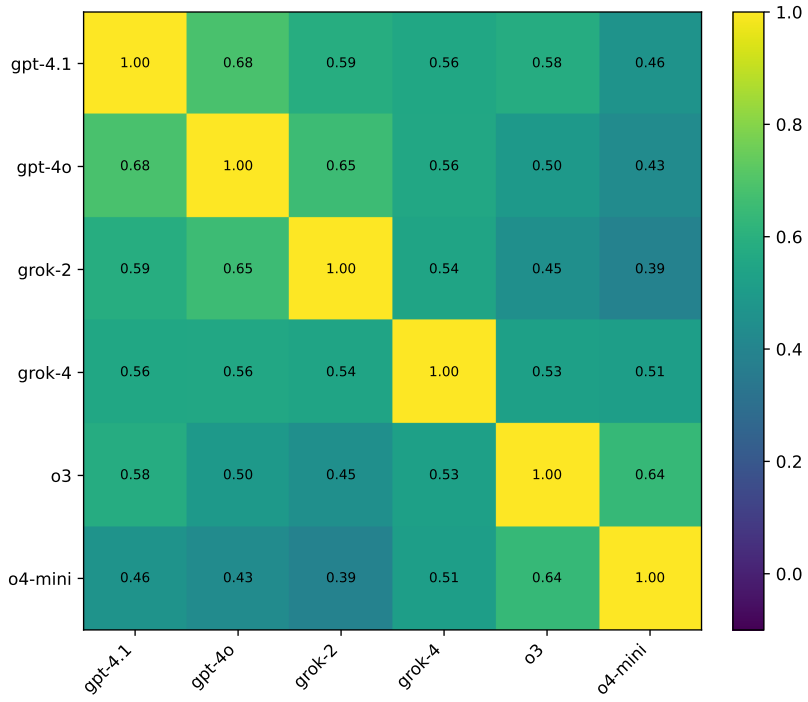


Figure 1: Agreement heatmap on *Required items with images* (majority-correct labels, $\geq 2/3$ runs).

"ultra-gold" benchmarks. For coverage context, Table 20 in Appendix G reports the distribution of items solved by exactly k models for both criteria (majority-correct and always-correct).

As a robustness view without chance correction, pairwise Jaccard overlaps on "solved" sets mirror the pattern (Appendix I): within-family overlaps are high (both families ≈ 0.63) and exceed cross-family (≈ 0.55), with the largest raw overlap for o3-o4-mini ($J \approx 0.72$).

4.5. Qualitative Error Analysis

We summarize the 83-item contrastive qualitative audit described in Section 3.4.4. Briefly, for each audited item we code one winner response (from the family that achieves $\geq 1/3$ correct runs) and one loser response (from the other family with $0/3$), using the merged taxonomy in Table 7; *No-error* is tracked separately and not counted as an error. Because the audit set is constructed to isolate family separation, reported proportions are descriptive of this contrastive set; we include 95% Wilson confidence intervals for key proportions (Appendix K.1).

Primary outcomes (Table 8) Visual misread was the dominant failure mode when errors occurred, including cases where the final answer happened to be correct. Because winner responses are selected from correct attempts by construction, *Visual misread* in the winner rows indicates "right answer for the wrong reason": this occurs in 15/68 (22.1%) of reasoning wins and 8/15 (53.3%) of non-reasoning wins (Table 8). Any response whose steps misinterpreted the figure was labeled *Visual misread* regardless of final-answer correctness. In the main stratum ($n = 68$), non-reasoning losers were labeled *Visual misread* in 58/68 (85.3%; 95% CI [75.0, 91.8]) and refused in 9/68 (13.2%; 95% CI [7.1, 23.3]), while reasoning winners were labeled *Visual misread* in 15/68 (22.1%; 95% CI [13.8, 33.3]) and *No-error* in 52/68 (76.5%; 95% CI

[65.1, 85.0]). In the counter-stratum ($n = 15$), reasoning losers refused in 8/15 (53.3%; 95% CI [30.1, 75.2]) and were labeled *Visual misread* in 7/15 (46.7%; 95% CI [24.8, 69.9]).

Explanation/format Explanation/format issues were rare in the main stratum and typically reflected a correct approach expressed in the wrong requested form (e.g., solving for x when the item asked for an expression).

Family differences The contrastive design isolates different failure modes by family: when non-reasoning models fail on items that reasoning models solve (main stratum), their errors are predominantly *Visual misread*; conversely, when reasoning models fail on items that non-reasoning models solve (counter stratum), failures split between refusal and visual misread. When reasoning succeeds in the main stratum, most responses are *No-error* with an adequate explanation (Table 8).

Table 7: Merged qualitative taxonomy used in the audit (three error categories). Each response receives one primary label; *No-error* (correct answer with an adequate explanation) is tracked separately and is not counted as an error.

Label	**Description**
Visual misread	• Misreading axes, tick spacing, or scale units • Failing to extract visual symbols (e.g., dots, blocks, shaded units, icons) • Misinterpreting geometric relations (e.g., parallel, similar, angles, tick marks) • Misreading embedded text or labels in the diagram
Explanation/format	• Correct or near-correct approach, but the explanation is terse, skips key steps, or is unclear • Correct reasoning in steps, but wrong final answer due to output format or misunderstanding of the requested form
Refused	`isSolvable=false` (model indicates the item is not solvable under the given inputs).

Table 8: Primary outcome distribution by problem set (merged taxonomy).

Stratum (family role)	Visual misread	Explanation/format	Refused	No-error
Counter (15) NON win	8/15 (53.3%)	4/15 (26.7%)	0/15 (0.0%)	3/15 (20.0%)
Counter (15) REASON lose	7/15 (46.7%)	0/15 (0.0%)	8/15 (53.3%)	0/15 (0.0%)
Main (68) REASON win	15/68 (22.1%)	1/68 (1.5%)	0/68 (0.0%)	52/68 (76.5%)
Main (68) NON lose	58/68 (85.3%)	1/68 (1.5%)	9/68 (13.2%)	0/68 (0.0%)

- Percentages over total problems in set; each row sums to 100%.
- *No-error* indicates a correct answer with an adequate explanation.
- *Refused* is an explicit decline to solve due to missing information.
- **NON** = non-reasoning model family; **REASON** = reasoning model family (Table 2).
- *win/lose* indicates whether that family solved the item under our run-level rule; the opposite family did not.
- Counter rows are the double-coded counter-stratum (disagreements adjudicated).
- Main rows are the single-coded main stratum.

5. Discussion

5.1. Answers to RQ1–RQ3

The quantitative results in Section 4 support three deployment-relevant conclusions: with images, model behavior is reasonably stable across repeated attempts, with tight uncertainty; when required figures are removed, models generally avoid speculative answering by refusing; and solvability judgments vary across models, motivating multi-model robustness views and our released benchmark subsets. Together, these findings argue for a reliability-first evaluation stance in educational settings: gate automated feedback and scoring not on single-run accuracy, but on within-model stability and cross-model corroboration.

5.2. Answer to RQ4 (qualitative patterns)

Across the contrastive audit, the primary family difference is *where* failures occur in the solution pipeline. When non-reasoning models fail on items that reasoning models solve, errors are dominated by *Visual misread* (Table 8), consistent with the idea that small perceptual or formatting differences can cascade into downstream symbolic errors (Closser et al., 2024). Reasoning-family responses show substantially fewer visual misreads in this stratum, aligning with prior evidence that stronger reasoning supervision can improve grounded, stepwise solutions (Wei et al., 2022; Kojima et al., 2022).

When non-reasoning models succeed (counter-stratum), residual errors more often reflect *Explanation/format* issues (e.g., underspecified steps or responding in the wrong requested form), which fits concerns about overconfident or incomplete rationales even when the final answer is correct (Kalai et al., 2025). Overall, the audit suggests a practical interpretation for tutoring: reasoning models more often clear the perceptual grounding step, while non-reasoning models' mistakes more frequently originate in visual evidence extraction; conversely, when ei-

ther family succeeds, the remaining gap is often communicative fidelity rather than core computation (Table 8).

5.3. Accessibility

The refusal-first behavior without images is a desirable safety property for assistive use: systems should elicit the missing figure (or a structured textual description) rather than speculate, consistent with guidance on relevance-preserving generative interactions in math tutoring (Levonian et al., 2025). Our released gold/ultra-gold subsets provide a practical testbed for accessibility: they are natural candidates for description-substitution experiments comparing original figures to vetted human (or model-assisted) descriptions while monitoring effects on accuracy, stability, and refusal.

5.4. Implications for ITS design, scoring, and learner modeling

For ITS deployment, the key question is not only whether a model is accurate, but whether its behavior is reliable enough to support *system design decisions*, automated scoring, and learner-model updates, consistent with prior educational LLM evaluation work that treats predictive consistency as a key dimension alongside accuracy (Tran et al., 2024). In image-dependent math, this reliability question is inseparable from *visual evidence availability*: when an item is REQUIRED, a system must distinguish learner error from missing-figure conditions before assigning credit or updating mastery. Our REQUIRED/USEFUL labeling and no-image refusal analysis operationalize this distinction, providing a practical criterion for when automated scoring and learner-model updates are valid versus when the system should first elicit the missing visual information. Our automated correctness pipeline (Appendix E.4) enables scalable solvability estimates, while run stability and cross-model agreement provide concrete *deployment gates*; this mirrors recent findings in automatic evaluation that panel-style corroboration across multiple LLM judges can improve reliability over single-judge settings (Verga et al., 2024; Badshah and Sajjad, 2025). For example, high-agreement items (e.g., those solved by ≥ 5 models at majority-correct; Appendix Table 20) are stronger candidates for automated feedback and downstream learner-model updates, since outcomes are less sensitive to sampling noise and model idiosyncrasies. Conversely, low-agreement items should trigger guardrails by design–fallback to human-authored hints, scaffolds that elicit explicit visual-evidence binding before algebra, or deferral–rather than treating model output as scored evidence. For deployment, no-image refusals can be treated as missing-context indicators and used to suppress automated scoring or learner-model updates until the required visual information is provided. When automated feedback is given, systems should prefer evidence-linked outputs that make clear what local signals support global judgments (Whitehill and LoCasale-Crouch, 2024).

6. Limitations

Our study has several limitations worth noting. We evaluated models on a single curriculum, which may not reflect the full diversity of middle-school math problems across formats, grades, or visual styles. The models themselves are evolving rapidly.

A related limitation is potential dependence on *image type* and diagram conventions. The items in our dataset include multiple visual forms (e.g., Cartesian graphs, geometric figures with conventional markings, measurement diagrams, and representational models such as tape

diagrams). Model performance may vary substantially across these types–for example, graphs and coordinate systems require accurate extraction of axes, scale, and plotted points; geometry relies on conventions such as tick marks, angle arcs, and parallel/perpendicular indicators; and real-world or schematic diagrams may require interpreting informal pictorial cues. Because we draw from a single curriculum, diagram styles and conventions are relatively consistent, which may make evaluation easier than settings where conventions vary across sources. Conversely, results may not transfer to curricula with different visual encodings, noisier renderings, or more heterogeneous diagram styles. Stratifying analyses by image type is an important next step for understanding when and why errors such as *Visual misread* occur and for designing targeted mitigations.

A further alternative explanation is potential *training-data exposure* (data contamination). IM materials are publicly available, so some item text (and possibly figures) could plausibly appear in web-scale corpora used to train or tune modern foundation models. If an evaluated model has memorized an item or a near-duplicate, it could produce a correct answer in the no-image condition without actually inferring missing visual information, inflating estimates of without-image solvability or "lucky-correct" behavior. This concern is consistent with evidence that some apparent mathematical "reasoning" in LLMs can be brittle under controlled perturbations (e.g., changing only numeric values) (Mirzadeh et al., 2025). Our design partially mitigates this concern by comparing with-image versus no-image behavior on the same prompt text and by manually auditing REQUIRED-labeled items that achieved any no-image correctness in the refinement pass (Appendix C), retaining items as REQUIRED when without-image success required assumptions or ambiguous inference. Nonetheless, we do not claim to fully rule out contamination, and dedicated checks (e.g., near-duplicate detection, paraphrase-based re-evaluation, or evaluation on newly authored items) remain important future work.

Image-role labeling necessarily involves judgment. The full 541-item labeling was produced by a single rater using a frozen operational codebook; to assess reliability, we conducted an independent second-rater audit on a stratified subset of 153 items and report chance-corrected agreement for both the four-way taxonomy and the REQUIRED vs. non-REQUIRED boundary (Appendix D). Consistent with prior observations that "usefulness" is more subjective, disagreements in the audit concentrated in the USEFUL vs. NOT REQUIRED distinction, while agreement on the high-stakes REQUIRED boundary was stronger.

To further reduce accidental over-labeling of REQUIRED, we performed a conservative flag-then-review refinement pass focused only on REQUIRED-labeled items, using no-image correctness under our prompt and scoring protocol only as a screening signal and adjudicating all final decisions against the operational definitions (Appendix C). This step strengthens confidence that items retained as REQUIRED contain task-critical information that is not recoverable from text alone without adding assumptions, but it may miss other boundary cases that different models, prompts, or scoring rules would surface; this dependence is likely conservative in that it tends to retain borderline items as REQUIRED rather than incorrectly downgrading them.

The qualitative audit, though focused and revealing, covered only 83 problems (a contrastive subset where the two model families separate under our run-level rule). Accordingly, we treat these codes as mechanism-finding and hypothesis-generating rather than statistically powered evidence for family-wide prevalence claims; scaling qualitative labeling to much larger item sets (e.g., with validated automated or LLM-assisted coding protocols) is an important direction for future work. We merged fine-grained visual errors (e.g., misreading axes vs. diagram relations) into a single *Visual misread* category due to rater disagreement–a necessary trade-off

for reliability, but one that obscures nuance. The original 7-category taxonomy is provided in Appendix L for transparency.

We did not include a human baseline–neither student nor teacher performance–so we cannot say how LLMs compare to learners in perception or explanation clarity. Nor did we explore alt-text generation, scaffolding prompts, or human-LLM collaboration, all promising directions for real-world tutoring.

A further limitation is that we do not analyze which textual cues in an item (e.g., explicit figure references such as "as shown in the diagram") predict refusal versus guessing in the without-image condition. The released item-level artifacts enable such analyses.

Finally, agreement metrics could be defined differently–for example, over non-refusal attempts or as a three-way code (correct/wrong/refusal). We chose correctness as the primary signal for pedagogical utility and report refusal separately (RQ2), but alternative definitions remain valuable robustness checks.

7. Future Work

Our findings highlight the superior visual perception and explanation clarity of reasoning models, alongside persistent gaps in non-reasoning models. To build on these insights and move toward practical, learner-facing systems, we outline several promising directions for future exploration.

A natural next step is to strengthen image-role labeling through multi-rater coding. A concise codebook could be applied by several raters, with agreement metrics and adjudication protocols reported. Re-running analyses on a consensus set, with sensitivity to labeling thresholds, would help assess the robustness of our Required classification.

To broaden generalizability, we plan to expand the corpus beyond Grades 6-8, while tagging items by figure type (e.g., graphs, diagrams, tables) and visual complexity. This would support strand-level analyses and difficulty modeling (e.g., using Rasch/IRT), treating model runs as responses.

For accessibility, the gold and ultra-gold benchmarks–items solved with images and responsibly refused without–offer an ideal testbed for textual description substitution. Randomized comparisons of vetted human descriptions versus original figures could measure effects on accuracy, stability, and refusal rates, with input from blind or low-vision users. A follow-up might explore whether model-generated alt text approaches human quality.

Explanation quality also merits deeper study. A teacher-developed rubric could score clarity, step coverage, and warranting in model rationales, correlating these with correctness and refusal behavior. This would clarify whether reasoning models truly excel pedagogically–or merely appear to–and whether scaffolds enhance clarity.

Finally, real-world pilots in intelligent tutoring systems remain a key goal: routing visual items to vetted models, inserting safety prompts when images are missing, and logging learner outcomes (e.g., hint needs, time-to-correct, refusal frequency). Ethical audits for hallucinated visual details and harmful failures would accompany deployment.

8. Conclusion

This paper examined whether contemporary MLLMs can solve middle-school mathematics problems whose figures carry essential information, and whether they behave responsibly when

those figures are absent. Using six models (three reasoning: grok-4, o3, o4-mini; three non-reasoning: gpt-4.1, gpt-4o, grok-2-vision), three runs per item in with-image and without-image conditions, and a conservative reclassification of image roles, we focused analyses on a final REQUIRED set of 376 items.

With images present, models showed non-trivial competence and high run stability across repeated attempts. Under our primary item-level metric (majority-correct, $\geq$2/3 runs), the strongest reasoning models achieved the highest solved rates (o4-mini 57.7%, o3 52.4%), while non-reasoning models ranged from the mid-30s to low-40s. Without images, models overwhelmingly refused ($\approx$95–99%) rather than guessing, with only rare lucky-correct answers ($\leq$3%). Agreement on which items are solvable with images was moderate overall (mean Cohen's $\kappa \approx 0.54$), with higher cross-model agreement within families than across families. We identified 76 items solved by all six models at majority-correct and 61 items solved 3/3 by all six, which we release as gold and ultra-gold benchmarks.

In a contrastive qualitative audit of 83 REQUIRED items (the exhaustive subset where the reasoning and non-reasoning families separate under our run-level criterion), visual misreading was the dominant failure mode. In the main stratum (reasoning succeeds; non-reasoning fails), non-reasoning *loser* responses were labeled *Visual misread* in 58/68 cases (85.3%), while reasoning *winner* responses were usually *No-error* (52/68; 76.5%) but still sometimes exhibited "right answer for the wrong reason" visual misreads (15/68; 22.1%). Explanation/format issues were rare in the main stratum, and were more evident among non-reasoning *winner* responses in the counter-stratum (4/15; 26.7%) (Table 8).

These findings have two immediate implications. For tutoring systems, image-role awareness should gate automated scoring and learner-model updates: when an item is REQUIRED, no-image refusals should be treated as missing-context signals rather than learner error, and with-image support should prioritize approaches that enforce explicit visual-evidence binding before algebra. For accessibility, the gold/ultra-gold benchmarks provide a principled starting point for description-substitution studies that test when structured textual descriptions can replace figures without degrading accuracy, stability, or refusal behavior.

Our study has limitations, including a single curricular source, evolving model versions, image-role labels that begin with a single rater (mitigated via an independent audit and a conservative refinement pass), and a focused contrastive qualitative audit. We did not evaluate alt-text generation, scaffolding prompts, or human–LLM collaboration.

We release code, prompts, summary tables, and the gold/ultra-gold benchmarks to support replication. Future work will strengthen multi-rater image-role reliability, stratify findings by figure type and visual complexity, test lightweight scaffolds targeting visual misreads and evidence binding, broaden coverage across grades and item types, and run accessibility studies comparing original figures to vetted textual descriptions. Together, these steps can turn measured competence on image-dependent items into more robust, learner-facing support.

ACKNOWLEDGMENTS

We would like to thank NSF (e.g., 2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), IES (e.g., R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, & R305A120125), GAANN (e.g., P200A180088 & P200A150306), EIR (U411B190024 & S411B210024), ONR (N00014-18-1-2768), NIH (R44GM146483), and Schmidt Futures. None of the opinions expressed here

are those of the funders.

Declaration of Generative AI Software Tools in the Writing Process

During the preparation of this work, the author(s) used Grok 4 (xAI) in the Qualitative Error Analysis (RQ4) section in order to iteratively refine the coding taxonomy, compute inter-rater reliability, interpret confusion matrices, and draft narrative interpretations of error patterns. The tool was also used to streamline and polish the Limitations and Future Work sections for clarity and flow. After using this tool, the author(s) reviewed and edited all content, verified all statistical claims and percentages against raw script output, and take full responsibility for the content of the publication.

References

Ainsworth, S. 2006. Deft: A conceptual framework for considering learning with multiple representations. *Learning and Instruction 16,* 3, 183–198.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., and Pelletier, R. 1995. Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences 4,* 2, 167–207.

Arneson, J. B. and Offerdahl, E. G. 2018. Visual literacy in Bloom: Using Bloom's taxonomy to support visual learning skills. *CBE–Life Sciences Education 17,* 1 (Mar), ar7.

ASSETS 2025 Organizing Committee. 2025. Accessibility guidelines. ACM SIGACCESS Conference on Computers and Accessibility. Accessed October 26, 2025.

Badshah, S. and Sajjad, H. 2025. Reference-guided verdict: LLMs-as-judges in automatic evaluation of free-form QA. In *Proceedings of the 9th Widening NLP Workshop*, C. Zhang, E. Allaway, H. Shen, L. Miculicich, Y. Li, M. M'hamdi, P. Limkonchotiwat, R. H. Bai, S. T.y.s.s., S. S. Han, S. Thapa, and W. B. Rim, Eds. Association for Computational Linguistics, Suzhou, China, 251–267.

Benetech. 2019. Specific guidelines – mathematics. http://diagramcenter.org/specific-guidelines-g.html. DIAGRAM Center, supported by the U.S. Department of Education, Office of Special Education Programs (Cooperative Agreement #H327B100001). Accessed October 26, 2025.

Braille Authority of North America (BANA). 2022. Guidelines and standards for tactile graphics. ISBN (print): 979-8-9883302-0-2; ISBN (braille): 979-8-9883302-1-9.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 16*, 321–357.

Chen, F., Yuan, H., Xu, Y., Feng, T., Cen, J., Liu, P., Huang, Z., and Yang, Y. 2025. MathFlow: Enhancing the perceptual flow of MLLMs for visual mathematical problems. arXiv preprint arXiv:2503.16549. Code available at https://github.com/MathFlow-zju/MathFlow.

Chen, Z., Hu, W., He, G., Deng, Z., Zhang, Z., and Hong, R. 2025. Unveiling uncertainty: A deep dive into calibration and performance of multimodal large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*. Association for Computational Linguistics, Abu Dhabi, UAE, 3095–3109.

Closser, A. H., Botelho, A., Chan, J., et al. 2024. Exploring the impact of symbol spacing and problem sequencing on arithmetic performance: An educational data mining approach. *Journal of Educational Data Mining 16,* 1, 84–111.

COBBE, K., KOSARAJU, V., BAVARIAN, M., CHEN, M., JUN, H., KAISER, L., PLAPPERT, M., TWOREK, J., HILTON, J., NAKANO, R., HESSE, C., AND SCHULMAN, J. 2021. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.

COHEN, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20,* 1, 37–46.

CORBETT, A. T. AND ANDERSON, J. R. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction 4,* 4, 253–278.

DIAGRAM CENTER. 2015. Image description guidelines. https://diagramcenter.org/making-images-accessible.html. Landing page with HTML and DOCX versions; accessed October 22, 2025.

EAGAN, B., BROHINSKY, J., WANG, J., AND SHAFFER, D. W. 2020. Testing the reliability of inter-rater reliability. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. LAK '20. Association for Computing Machinery, New York, NY, USA, 454–461.

EVAGOROU, M., ERDURAN, S., AND MÄNTYLÄ, T. 2015. The role of visual representations in scientific practices: From conceptual understanding and knowledge generation to 'seeing' how science works. *International Journal of STEM Education 2,* 11 (July), 11.

FENG, J., WANG, Z., ZHANG, Z., GUO, Y., ZHOU, Z., CHEN, X., LI, Z., AND YIN, D. 2025. MathReal: We keep it real! A real scene benchmark for evaluating math reasoning in multimodal large language models. arXiv preprint arXiv:2508.06009. Code available at https://github.com/junfeng0288/MathReal.

HEFFERNAN, N. T. AND HEFFERNAN, C. L. 2014. The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *International Journal of Artificial Intelligence in Education 24,* 4, 470–497.

HWANG, H., KIM, D., KIM, S., YE, S., AND SEO, M. 2024. Self-explore: Enhancing mathematical reasoning in language models with fine-grained rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Association for Computational Linguistics, Miami, Florida, USA, 1444–1466.

JI, H., QIU, S., XIN, S., HAN, S., CHEN, Z., ZHANG, D., WANG, H., AND YAO, H. 2025. From EduVisBench to EduVisAgent: A benchmark and multi-agent framework for reasoning-driven pedagogical visualization. In *The 5th Workshop on Mathematical Reasoning and AI at NeurIPS 2025*. NeurIPS, NeurIPS, Online.

JIANG, Z., PENG, H., FENG, S., LI, F., AND LI, D. 2024. LLMs can find mathematical reasoning mistakes by pedagogical chain-of-thought. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, K. Larson, Ed. International Joint Conferences on Artificial Intelligence Organization, Jeju Island, South Korea, 3439–3447. Main Track.

KADAVATH, S., CONERLY, T., ASKELL, A., EL-SHOWK, S., SCHIEFER, N., NADLER, K., LADD, A., GANGULI, D., HENIGHAN, T., JONES, A., BOWMAN, N., KRAVEC, A., LOVITT, Z., NDOUSSE, K., CHEN, A., KAPADIA, T., AMODEI, D., HERNANDEZ, D., DRAIN, D., GANGULI, S., CLARK, J., AND KAPLAN, J. 2022. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221.

KALAI, A. T., NACHUM, O., VEMPALA, S. S., AND ZHANG, E. 2025. Why language models hallucinate. arXiv preprint arXiv:2509.04664.

KOEDINGER, K. R. AND CORBETT, A. T. 2006. Cognitive tutors: Technology bringing learning science to the classroom. In *The Cambridge Handbook of the Learning Sciences*, R. K. Sawyer, Ed. Cambridge University Press, New York, NY, 61–77.

KOJIMA, T., GU, S. S., REID, M., MATSUO, Y., AND IWASAWA, Y. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems 35*, 22199–22213.

KRIPPENDORFF, K. 2011. Computing Krippendorff's Alpha-Reliability. Working Paper 43, University of Pennsylvania, Annenberg School for Communication, Philadelphia, PA. Jan. Postprint version.

LARKIN, J. H. AND SIMON, H. A. 1987. Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science 11,* 1, 65–100.

LEVONIAN, Z., HENKEL, O., LI, C., AND POSTLE, M.-E. 2025. Designing safe and relevant generative chats for math learning in intelligent tutoring systems. *Journal of Educational Data Mining 17,* 1, 66–97.

LI, C., LIU, Y., ZHANG, T., WANG, M., AND HUANG, H. 2026. VisioMath: Benchmarking figure-based mathematical reasoning in LMMs. ICLR 2026 Poster, OpenReview. https://openreview.net/forum?id=31pzK2VSQQ.

LU, P., BANSAL, H., XIA, T., LIU, J., LI, C., HAJISHIRZI, H., CHENG, H., CHANG, K.-W., GALLEY, M., AND GAO, J. 2024. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, Vienna, Austria.

MARRIOTT, K., LEE, B., BUTLER, M., CUTRELL, E., ELLIS, K., GONCU, C., HEARST, M., MCCOY, K., AND SZAFIR, D. A. 2021. Inclusive data visualization for people with disabilities: A call to action. *Interactions 28,* 3 (Apr.), 47-51.

MATHEMATICS, I. 2019. Illustrative mathematics, grade 6-8. Available at https://illustrativemathematics.org/. Authored by Illustrative Mathematics.

MAYER, R. E. 2020. *Multimedia Learning*, 3rd ed. Cambridge University Press, Cambridge, UK.

MIRZADEH, S. I., ALIZADEH, K., SHAHROKHI, H., TUZEL, O., BENGIO, S., AND FARAJTABAR, M. 2025. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*. OpenReview.net, Singapore.

NOBRE, C., ZHU, K., MÖRTH, E., PFISTER, H., AND BEYER, J. 2024. Reading between the pixels: Investigating the barriers to visualization literacy. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. CHI '24. Association for Computing Machinery, New York, NY, USA.

NYE, M., ANDREASSEN, A. J., GUR-ARI, G., MICHALEWSKI, H., AUSTIN, J., BIEBER, D., DOHAN, D., LEWKOWYCZ, A., BOSMA, M., LUAN, D., SUTTON, C., AND ODENA, A. 2021. Show your work: Scratchpads for intermediate computation with language models. arXiv preprint arXiv:2112.00114.

OUYANG, L., WU, J., JIANG, X., ALMEIDA, D., WAINWRIGHT, C., MISHKIN, P., ZHANG, C., AGARWAL, S., SLAMA, K., RAY, A., ET AL. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems 35*, 27730–27744.

PELÁNEK, R. 2017. Bayesian knowledge tracing, logistic models, and beyond: An overview of learner modeling techniques. *User Modeling and User-Adapted Interaction 27,* 3, 313–350.

ROSCHELLE, J., FENG, M., MURPHY, R. F., AND MASON, C. A. 2016. Online mathematics homework increases student achievement. *AERA Open 2,* 4, 1–12.

RUDMAN, W., GOLOVANEVSKY, M., BAR, A., PALIT, V., LECUN, Y., EICKHOFF, C., AND SINGH, R. 2025. Forgotten polygons: Multimodal large language models are shape-blind. In *Findings of the Association for Computational Linguistics: ACL 2025*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Association for Computational Linguistics, Vienna, Austria, 11983–11998.

SELENT, D., PATIKORN, T., AND HEFFERNAN, N. T. 2016. ASSISTments dataset from multiple randomized controlled experiments. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*. L@S '16. Association for Computing Machinery, Edinburgh, Scotland, UK, 181–184.

SHI, W., HU, Z., BIN, Y., LIU, J., YANG, Y., NG, S.-K., BING, L., AND LEE, R. K.-W. 2024. Math-LLaVA: Bootstrapping mathematical reasoning for multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, Miami, Florida, USA, 4663–4680.

SHI, Y., LIANG, R., AND XU, Y. 2025. EducationQ: Evaluating LLMs' teaching capabilities through multi-agent dialogue framework. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Association for Computational Linguistics, Vienna, Austria, 32799–32828.

SUN, Y., ZHANG, S., TANG, W., CHEN, A., KONIUSZ, P., ZOU, K., XUE, Y., AND VAN DEN HENGEL, A. 2026. Math blind: Failures in diagram understanding undermine reasoning in MLLMs. ICLR 2026 Poster, OpenReview. https://openreview.net/forum?id=RtvmTxdQV9.

SWELLER, J., VAN MERRIËNBOER, J. J. G., AND PAAS, F. G. W. C. 1998. Cognitive architecture and instructional design. *Educational Psychology Review 10,* 3, 251–296.

SÁEZ, A. I., RHOMRASI, L., AHSINI, Y., VINUESA, R., HOYAS, S., SABATER, J. P. G., I ALFONSO, M. J. F., AND CONEJERO, J. A. 2025. Evaluating visual mathematics in multimodal LLMs: A multilingual benchmark based on the kangaroo tests. arXiv preprint arXiv:2506.07418.

TANG, L., KIM, G., ZHAO, X., LAKE, T., DING, W., YIN, F., SINGHAL, P., WADHWA, M., LIU, Z. L., SPRAGUE, Z., NAMUDURI, R., HU, B., RODRIGUEZ, J. D., PENG, P., AND DURRETT, G. 2025. ChartMuseum: Testing visual reasoning capabilities of large vision-language models. In *Advances in Neural Information Processing Systems 38*. Curran Associates, Inc., San Diego, CA.

TRAN, N., PIERCE, B., LITMAN, D., CORRENTI, R., MATSUMURA, L. C., ET AL. 2024. Multi-dimensional performance analysis of large language models for classroom discussion assessment. *Journal of Educational Data Mining 16,* 2, 304–335.

TREWIN, S. 2019. Describing figures. SIGACCESS Resources. Accessed October 26, 2025.

VANLEHN, K. 2006. The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education 16,* 3, 227–265.

VANLEHN, K. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist 46,* 4, 197–221.

VERGA, P., HOFSTÄTTER, S., ALTHAMMER, S., SU, Y., PIKTUS, A., ARKHANGORODSKY, A., XU, M., WHITE, N., AND LEWIS, P. 2024. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. arXiv preprint arXiv:2404.18796.

W3C ACCESSIBILITY GUIDELINES WORKING GROUP. 2024. Web content accessibility guidelines (wcag) 2.2. W3C Recommendation. Latest version: https://www.w3.org/TR/WCAG22/.

W3C WEB ACCESSIBILITY INITIATIVE. 2022. Images tutorial: Complex images. https://www.w3.org/WAI/tutorials/images/complex/. Updated 17 January 2022. Accessed 24 Oct 2025.

WANG, P., LI, Z.-Z., YIN, F., RAN, D., AND LIU, C.-L. 2025. MV-MATH: Evaluating multimodal math reasoning in multi-visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Nashville, TN, USA, 19541–19551.

WEI, J., WANG, X., SCHUURMANS, D., BOSMA, M., ICHTER, B., XIA, F., CHI, E. H., LE, Q. V., AND ZHOU, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances*

in Neural Information Processing Systems 35. Curran Associates, Inc., New Orleans, Louisiana, USA.

WHITEHILL, J. AND LOCASALE-CROUCH, J. 2024. Automated evaluation of classroom instructional support with LLMs and bows: Connecting global predictions to specific feedback. *Journal of Educational Data Mining 16,* 1, 34–60.

WOOLF, B. P. 2008. *Building Intelligent Interactive Tutors: Student-Centered Strategies for Revolutionizing E-Learning*. Morgan Kaufmann, San Francisco, CA.

XIA, S., LI, X., LIU, Y., WU, T., AND LIU, P. 2025. Evaluating mathematical reasoning beyond accuracy. *Proceedings of the AAAI Conference on Artificial Intelligence 39,* 26, 27723–27730.

YAN, Y., SU, J., HE, J., FU, F., ZHENG, X., LYU, Y., WANG, K., WANG, S., WEN, Q., AND HU, X. 2025. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, Vienna, Austria, 11798–11827.

YAN, Y., WANG, S., HUO, J., LI, H., LI, B., SU, J., GAO, X., ZHANG, Y.-F., XU, T., CHU, Z., ZHONG, A., WANG, K., XIONG, H., YU, P. S., HU, X., AND WEN, Q. 2024. ErrorRadar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. Submitted to ICLR 2025; OpenReview: https://openreview.net/forum?id=GeTBk67mK6.

YANG, Z., LI, L., LIN, K., WANG, J., LIN, C.-C., LIU, Z., AND WANG, L. 2023. The dawn of LMMs: Preliminary explorations with GPT-4V(ision). arXiv preprint arXiv:2309.17421.

YIN, S., FU, C., ZHAO, S., LI, K., SUN, X., XU, T., AND CHEN, E. 2024. A survey on multimodal large language models. *National Science Review 11,* 12 (Nov.), nwae403.

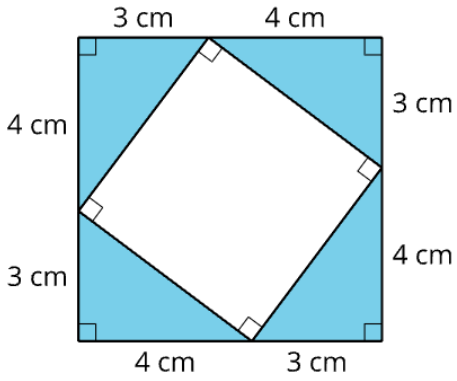
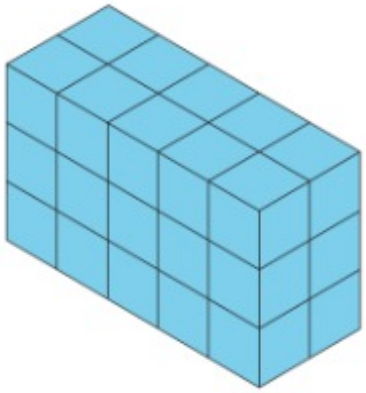
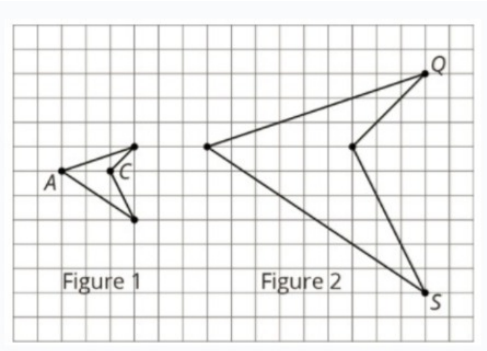
YIN, Z., SUN, Y., HUANG, X., QIU, X., AND ZHAO, H. 2025. Error classification of large language models on math word problems: A dynamically adaptive framework. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics, Suzhou, China, 338–365.

ZHANG, L. AND GRAF, E. A. 2025. Mathematical computation and reasoning errors by large language models. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers*. Vol. 2025.aimecon-main. National Council on Measurement in Education (NCME), Pittsburgh, Pennsylvania, United States, 417–424.

ZHANG, R., JIANG, D., ZHANG, Y., LIN, H., GUO, Z., QIU, P., ZHOU, A., LU, P., CHANG, K.-W., QIAO, Y., ET AL. 2024. MATHVERSE: Does your multi-modal LLM truly see the diagrams in visual math problems? In *European Conference on Computer Vision*. Springer Nature Switzerland, Cham, 169–186.

A. IMAGE-ROLE EXEMPLARS

Table 9: Image-role exemplars, with item GUIDs included for traceability.

Required: Task-critical information appears only in the image. *Item ID:* `3b119229-571b-4dfc-be97-d88c62a41a11`  Find the area of the shaded region(s) of the figure. _____ cm^2	**Useful:** Text sufficient; image clarifies/disambiguates. *Item ID:* `621e3968-48d9-44a7-a36c-c120e25682bd`  A rectangular prism is 3 units high, 2 units wide, and 5 units long. What is its surface area in square units? _____ square units
Not Required: Decorative / no task-relevant support. *Item ID:* `d7b5f135-04ff-419b-a005-cf6e4a7d5851`  "Neon Bracelets" by Public Domain. CC0. Neon bracelets cost $1 for 4. What is the cost per bracelet? $_____	**Insufficient:** Missing information prevents a unique answer. *Item ID:* `0f88aaba-bb92-4b0f-9252-b104f81ba8fb`  What is the distance between B and D? _____ units

B. MODEL-RESPONSE EXEMPLARS

Note. Exemplars are abridged for readability. Each table shows (left) the item stem, item GUID, and figure, and (right) two response excerpts (one non-reasoning and one reasoning) selected from the contrastive audit described in Section 3. Excerpts follow the requested JSON schema (showing `isSolvable`, a minimal step summary, and the extracted `answer.value`/`answer.units`). The qualitative label shown is the paper's primary merged code for that response (Table 7); we do not re-adjudicate codes based on excerpted details. The ✓/✗indicator denotes whether the extracted final answer matches the reference answer under our automated equivalence procedure (Appendix E.4).

Table 10: **Model-response exemplar 1 (paired; area from grid).** Both responses apply the correct area formula, but the reasoning-family response misreads the figure's height, producing an incorrect final answer (✗).

Problem (with image)	**Model response excerpt (JSON schema)**

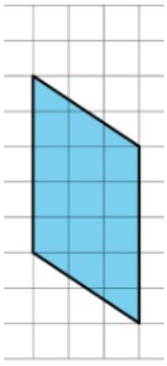
Item: `298a3eb5-21fb-4923-b7df-2cd6200d9ec`
Stem: Find the area of the parallelogram.
Gold (answer): 15
 
	Non-reasoning (gpt-4o; Attempt 1) — *No-error*
	✓ `isSolvable: true`
	`Step 1: Base length is 3 units.`
	`Step 2: Height is 5 units.`
	`Step 3: Area = base × height.`
	`answer.value:` **`15`**
	`answer.units: square units`
	Reasoning (o4-mini-2025-04-16; Attempt 1) — *Visual misread*
	✗ `isSolvable: true`
	`Step 1: Base spans 3 squares.`
	`Step 2: Height spans 7 squares.`
	`Step 3: Area = base × height.`
	`answer.value:` **`21`**
	`answer.units: square units`

Table 11: **Model-response exemplar 2 (paired; correctness vs. explanation fidelity).** The non-reasoning response attains the correct final answer (✓) but is coded *Explanation/format* because its stated visual grounding is inconsistent with the figure; the reasoning-family response misreads the diagram and yields an incorrect value (✗).

Problem (with image)	**Model response excerpt (JSON schema)**

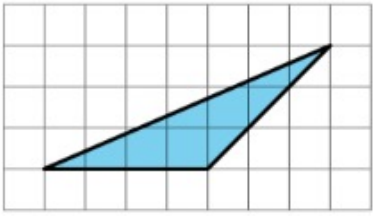
Item: `4b338ede-f1f4-4fb7-903b-93788c3dedce`
Stem: Find the area of the triangle.
Gold (answer): 6
 
	Non-reasoning (grok-2-vision-1212; Attempt 3) — *Explanation/format*
	✓ `isSolvable: true`
	`Step 1: Identify triangle vertices from the grid.`
	`Step 2: Use a coordinate-based triangle area formula.`
	`Step 3: Substitute and simplify.`
	`answer.value: 6`
	`answer.units: square units`
	Reasoning (grok-4-0709; Attempt 1) — *Visual misread*
	✗ `isSolvable: true`
	`Step 1: Identify triangle vertices from the grid.`
	`Step 2: Treat as a right triangle using base/height from the grid.`
	`Step 3: Area` $= \frac{1}{2}bh$.
	`answer.value: 7.5`
	`answer.units: square units`

Table 12: **Model-response exemplar 3 (paired; axes/interval selection).** The non-reasoning response selects the wrong interval on the graph and computes an incorrect rate (✗); the reasoning-family response selects the correct decreasing interval and computes the correct rate magnitude (✓).

Problem (with image)	**Model response excerpt (JSON schema)**

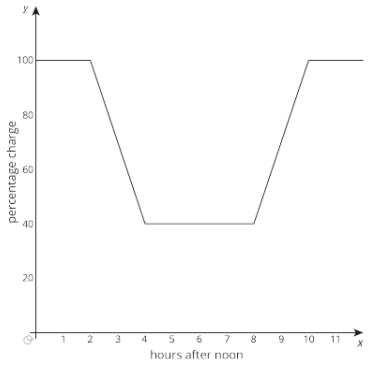
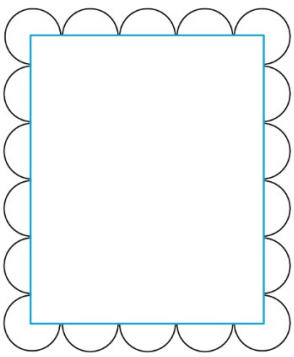
Item: `2a668c96-77b3-4c2c-b199-8396f2f13372`
Stem: Determine the phone battery discharge rate during the interval where it is decreasing.
Gold (answer): 30
 
	Non-reasoning (gpt-4o; Attempt 1) — *Visual misread*
	✗ `isSolvable: true`
	`Step 1: Use interval 1–4 hours after noon.`
	`Step 2: Read 100% at 1 hour and 40% at 4 hours.`
	`Step 3: Change` $= 100 - 40 = 60$`, duration` $= 3$`.`
	`Step 4: Rate` $= 60/3 = 20$`.`
	`answer.value:` **`20`**
	`answer.units: % per hour`
	Reasoning (o4-mini-2025-04-16; Attempt 2) — *No-error*
	✓ `isSolvable: true`
	`Step 1: Use interval 2–4 hours after noon (100% to 40%).`
	`Step 2: Change` $= 40 - 100 = -60$`, duration` $= 2$`.`
	`Step 3: Rate` $= (-60)/2 = -30$`, magnitude 30 per hour.`
	`answer.value:` **`30`**
	`answer.units: % per hour`

Table 13: **Model-response exemplar 4 (without-image; RQ2).** Under image removal, one model proceeds as if the task were a plain rectangle-perimeter computation (incorrect), whereas another model refuses and requests the missing visual information. The figure is shown here for reader context but was withheld from the models in the *without-image* condition. Markers: ✓ correct final answer; ✗ incorrect final answer; ⊘ refusal (`isSolvable=false`).

Problem (without-image condition)	**Model response excerpt (JSON schema)**

Item: `b65e19a7-b903-478c-af5a-a0aab98f689a`
Problem ID: `PRBE4YB`
Stem: Kiran bent some wire around a rectangle to make a picture frame (8 in by 10 in). Find the perimeter of the wire picture frame (nearest tenth).
Gold (answer): 62.8
Condition: *without-image* (models did not receive the figure).
 
	o3-2025-04-16 (Attempt 3) — *Guess-wrong (no-image)*
	✗ `isSolvable: true`
	`Step 1: Use rectangle dimensions 10 and 8.`
	`Step 2: Perimeter` $= 2(10 + 8) = 36$.
	`Step 3: Round to nearest tenth.`
	`answer.value: 36.0`
	`answer.units: inches`
	grok-4-0709 (Attempt 3) — *Refused (no-image)*
	⊘ `isSolvable: false`
	`required_information: [need the figure / a description of the wire path]`
	`answer: null`

C. REFINEMENT AUDIT TRAIL (REQUIRED → USEFUL)

C.1. AUDIT-SET COMPOSITION FOR INTER-RATER RELIABILITY

The stratified IRR audit set ($n = 153$) included all items labeled NOT REQUIRED ($n = 26$) and INSUFFICIENT ($n = 17$), all refinement-boundary items (the six items downgraded REQUIRED→USEFUL and the 14 items flagged for review but retained as REQUIRED), and additional seeded random samples of REQUIRED and USEFUL items to reach targets of 70 and 40, respectively.

C.2. SCREENING SIGNAL

To audit potential over-labeling of REQUIRED items, we manually reviewed every *Required*-labeled item that received at least one correct answer in the no-image condition (i.e., the model answered correctly despite the figure being withheld). No-image correctness was used only to prioritize items for inspection; it was not used as a decision rule for re-labeling. Each flagged item was adjudicated against the operational definitions of image role: if the text alone fully specified a unique solution without introducing additional assumptions, the item was downgraded REQUIRED→USEFUL; otherwise it remained REQUIRED and was recorded as assumption/ambiguity-driven without-image success.

In the initial screening (using gpt-4o and grok-2), 17 items were flagged; 6 were downgraded and 11 were retained as REQUIRED due to without-image ambiguity. After solving with all six models, three additional items were flagged (correct in no-image only by reasoning models); manual inspection retained all three as REQUIRED.

Table 14: Items downgraded from REQUIRED to USEFUL after manual inspection. No-image correctness served only as a screening signal to prioritize inspection; downgrade decisions required that the text alone support a unique solution without extra assumptions.

Item GUID	**Manual adjudication (without-image; no assumptions)**
`abdfafc7-9963-402e-945f-77ea3e6e2dbe`	All needed quantities are stated in text (1146 in^2, 10 in, 14 bags); compute $(1146 \cdot 10)/14 \approx 819\ \text{in}^3$.
`254bf2a3-5b60-4b66-8870-f2b6ee9aaa8b`	Andre's savings are fully specified in text ($15 initial, $5/week); equation $y = 5x + 15$; Noah's graph not required.
`b0c42988-6a6a-4cb0-b206-19e6beda4c44`	Given diameter $d = 2$ cm, radius $r = d/2 = 1$ cm; the graph/context is irrelevant to this subquestion.
`7c0ce88e-a06a-488e-9dca-31ce626792ea`	A dilation about B maps H to itself only with identity scale factor $k = 1$ (for $H \neq B$); no measurements required.
`6d57b7ce-512d-4d09-8234-cc119b87429a`	Two points are given in text, $(0, 2)$ and $(4, 1)$; the line through them is $y = -\frac{1}{4}x + 2$; no image-based information needed.
`c4949d14-d7ab-4d3d-ac27-44dc709a710b`	Volume $V = Ah$; doubling h doubles V for any base area A, so percent increase is 100%.

Table 15: Items flagged for review due to at least one correct no-image run (initially on gpt-4o/grok-2, and again after adding four more models), but retained as REQUIRED after manual inspection because without-image success depended on assumptions, ambiguous diagram interpretation, or unstated scale/structure.

Item GUID	Without-image success requires inference (retained REQUIRED)
7abddce0-3a22-4462-b584-7027be922a40	Tape diagram segmentation (4 and 5 equal sub-rectangles).
aa2aca39-5f91-45cd-b54f-a74710028c5b	Interpretation of $\overline{CD}$ (diameter vs. chord).
20b68591-2406-4fc6-829a-048825657193	Rectangle dimensions (3×5 inferred from figure).
fc23d2c5-3fd2-44e9-88e1-82fee78641e1	Segment overlap interpretation ($\overline{BC}$ vs. $\overline{BE}$).
17580d14-c356-47f2-b586-a095e5bd649b	Midpoint inference on a 0-6 number line.
495db94a-bca6-421c-9f18-0c9439d7676f	Partition inference: $\overline{AF}$ and $\overline{FD}$ treated as a non-overlapping partition of $\overline{AD}$.
d69b0a8d-4d35-4d21-9d64-18c5f80d950b	Point-to-line assignment ($C, D \in k$; $A, B \in l$).
708be5ac-f27b-4d25-8bc1-436673265ba9	Route-geometry inference: segments treated as straight and arranged as shown.
4816db56-78f0-40b4-bcf7-23e067ed6964	Shape classification inference (rectangle interpreted as non-square).
dd1cafca-786f-4dca-b816-74fe3762c4e6	Grid-structure inference: cube face interpreted as a 3×3 grid.
18b5ee97-7592-4e7e-b8bc-4e824033ebe3	Tick-mark count/spacing inference between 6 and 15.
1673fe10-8161-4479-9d0c-3df4ba9f55cb	Box-plot structure inference: assumes 0.5 and 1 are Q1 and Q3 (the box).
96c844c5-ef04-4945-82c5-5e4fb5047cae	Diagram-dependent configuration; overlap area non-unique from text alone.
47f829d6-62ce-4880-9182-f35006db5acf	Angle-label inference: b treated as $m\angle ABC$, with $DE \parallel AC$.

D. INTER-RATER RELIABILITY FOR IMAGE-ROLE LABELS

We report confusion matrices and chance-corrected agreement (Cohen's κ) for the independent image-role labels on the 153-item IRR audit set (Section 3). We additionally report 95% bootstrap confidence intervals (items resampled with replacement; $B = 5000$; seed=123). Across the four labels (REQUIRED, USEFUL, NOT REQUIRED, INSUFFICIENT), agreement was 71.9% (110/153) with $\kappa = 0.591$ (95% CI [0.492, 0.687]). When collapsing to a binary boundary (REQUIRED vs. non-REQUIRED), agreement was 88.2% (135/153) with $\kappa = 0.764$ (95% CI [0.657, 0.865]).

Table 16: Confusion matrix for four-way image-role labels on the 153-item IRR audit set (rows: independent rater; columns: first author).

	Rater 2 (first author)			
Rater 1	REQUIRED	USEFUL	NOT REQUIRED	INSUFFICIENT
REQUIRED	62	5	0	5
USEFUL	1	21	4	1
NOT REQUIRED	1	13	16	0
INSUFFICIENT	6	1	6	11

Table 17: Confusion matrix for the collapsed boundary (REQUIRED vs. non-REQUIRED), where NOT REQUIRED aggregates USEFUL, NOT REQUIRED, and INSUFFICIENT.

	Rater 2 (first author)	
Rater 1	REQUIRED	non-REQUIRED
REQUIRED	62	10
non-REQUIRED	8	73

E. PROMPTS, RUBRICS, AND NORMALIZATION

E.1. SETUP

Each Required item was evaluated under two conditions: With images (problem text plus its figure) and Without images (problem text only; any HTML `<img>` tags anonymized). For the image condition, the figure was included in the request payload as base64-encoded image data. Every item was run *three* times per model. The exact same prompt template was used for all six models.

E.2. PROMPT TEMPLATE AND EXPECTED JSON

Models were instructed to return a step-by-step solution in a fixed JSON structure.

Problem Statement: "`<question_text>`"

Instruction: "Provide a step-by-step solution in JSON format."

Response Structure:

```
{
  "solution": {
    "isSolvable": true,
    "required_information": [],
    "steps": [
      {"step_number": 1, "description": "First step."},
      {"step_number": 2, "description": "Next step."}
    ],
    "answer": {"value": "<answer_value>", "units": "<units>"}
  }
}
```

If the model could not solve the problem due to missing information, it was instructed to refuse and to list the required missing information:

```
{
  "solution": {
    "isSolvable": false,
    "required_information": ["<missing_info_1>", "<missing_info_2>"],
    "steps": [],
    "answer": null
  }
}
```

E.3. Rubrics: Scoring and Aggregation

E.3.1. Run-level scoring.

A run is marked correct if the extracted final answer `answer.value` matches the reference answer (unitless in this filtered dataset), using exact string match and, when needed, symbolic equivalence for numeric/algebraic forms. The `answer.units` field is recorded but does not affect correctness scoring. A run is marked as a refusal only when the returned JSON sets `isSolvable=false`. Any `required_information` the model provides is recorded but does not affect scoring.

E.3.2. Item-level aggregation

Per item and model, we count the number of correct runs across the three attempts:

$$3/3,\ 2/3,\ 1/3,\ 0/3.$$

We use these to form the sets:

- **Always** (3/3): items solved correctly in all three runs.
- **Majority** ($\geq$ 2/3): items solved correctly in at least two of three runs (i.e., 3/3 or 2/3).

"Consistency" at the item level refers to whether the *correctness label* (correct vs. incorrect) is identical across the three runs for a given model.

E.4. Answer normalization & equivalence

We compare the model's final answer string to the gold answer using a two-stage procedure: lightweight text normalization and parsing with SymPy; algebraic equivalence by symbolic simplification.

E.4.1. Preprocessing

Before parsing, we trim whitespace and drop a trailing degree symbol from the prediction (e.g., "30°" $\rightarrow$ "30").

E.4.2. Normalization rules

Given an answer string s, we apply:

1. **Right-hand side of equations:** if s contains "=", keep the substring after the last "=" (e.g., "`x=5`" $\rightarrow$ "`5`").
2. **Implicit multiplication:** insert "`*`" where omitted: digit-letter ("`3x`" $\rightarrow$ "`3*x`"), close-paren-letter ("`(x+1)y`" $\rightarrow$ "`(x+1)*y`"), and digit-"`sqrt`"/letter ("`3sqrt(10)`" $\rightarrow$ "`3*sqrt(10)`").
3. **Negative groups:** replace "`(-`" by "`(-1*`" (e.g., "`(-x+2)`").
4. **Mixed numbers:** strings of the form "`a b/c`" are mapped to $a + \frac{b}{c}$ (e.g., "`3 3/4`" $\rightarrow \frac{15}{4}$).
5. **Parsing:** parse with SymPy (`parse_expr`, standard transformations, `evaluate=False`) using a local symbol table for `x`, `y`, and `sqrt`.

E.4.3. Equivalence test

We first check exact string equality. Otherwise, let $\hat{y}$ and y^* be the parsed prediction and gold expressions; we declare equivalence when

$$\text{simplify}\big(\,\hat{y} - y^*\,\big) \;=\; 0.$$

E.4.4. Scope

This procedure targets common formatting differences (implicit multiplication, mixed numbers, degree mark). It does not apply numeric tolerances or unit conversions, and it does not attempt recovery from parsing failures or more exotic symbolic forms.

F. Run-level and Item-level Tallies (Required, with images)

Table 18: Run-level tallies on *Required, with images* (counts and percentages over 1,128 runs per model). Codes refer to Table 2.

	Correct		**Unsolvable**		**Wrong**	
Model	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
Non-reasoning						
gpt-4.1	482	42.7%	176	15.6%	470	41.7%
gpt-4o	425	37.7%	150	13.3%	553	49.0%
grok-2	389	34.5%	75	6.7%	664	58.9%
Reasoning						
grok-4	458	40.6%	357	31.6%	313	27.8%
o3	604	53.6%	215	19.1%	309	27.4%
o4-mini	650	57.6%	121	10.7%	357	31.6%

Table 19: Required with images: item-level outcomes across three runs per model. Entries are counts with percent of the (N=376) Required items; columns report items solved in all three runs (3/3), exactly two (2/3), exactly one (1/3), and none (0/3). Model codes follow Table 2.

	3/3		2/3		1/3		0/3	
Model	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
Non-reasoning								
gpt-4.1	130	34.6%	30	8.0%	32	8.5%	184	48.9%
gpt-4o	113	30.1%	20	5.3%	46	12.2%	197	52.4%
grok-2	101	26.9%	27	7.2%	32	8.5%	216	57.4%
Reasoning								
grok-4	126	33.5%	25	6.7%	30	8.0%	195	51.9%
o3	164	43.6%	33	8.8%	46	12.2%	133	35.4%
o4-mini	177	47.1%	40	10.6%	39	10.4%	120	31.9%

G. Items Solved by Exactly k Models (Required, with images)

Table 20: Required + image: items solved by exactly k of the six models under two criteria. Values are counts and percentages of the $N = 376$ Required items. Criteria: Majority ($\geq$ 2/3 runs) and Always (3/3 runs). Δ is the difference in percentage points (Always $-$ Majority).

k	**Majority ($\geq$2/3)**		**Always (3/3)**		Δ **pp**
	Count	**Percent**	**Count**	**Percent**	
0	113	30.1%	156	41.5%	+11.4
1	49	13.0%	40	10.6%	-2.4
2	30	8.0%	32	8.5%	+0.5
3	42	11.2%	34	9.0%	-2.1
4	35	9.3%	26	6.9%	-2.4
5	31	8.2%	27	7.2%	-1.1
6	76	20.2%	61	16.2%	-4.0
Total	376	100.0%	376	100.0%	0.0

H. Pairwise Contingency Counts for Cohen's κ

Table 21: Pairwise contingency counts for Cohen's κ (Required, with images; N=376 items). TP: both solved (majority-correct, $\geq$2/3 runs); FN: Row solved/Column not; FP: Row not/Column solved; TN: both not solved. Full matrices for all 15 pairs available as supplementary CSV.

Model Pair	**TP**	**FN**	**FP**	**TN**
gpt-4.1 vs. gpt-4o	118	42	15	201
gpt-4.1 vs. grok-2	107	53	21	195
gpt-4o vs. grok-2	101	32	27	216
gpt-4.1 vs. grok-4	115	45	36	180
gpt-4o vs. grok-4	103	30	48	195
grok-2 vs. grok-4	99	29	52	196
gpt-4.1 vs. o3	139	21	58	158
gpt-4o vs. o3	117	16	80	163
grok-2 vs. o3	110	18	87	161
gpt-4.1 vs. o4-mini	137	23	80	136
gpt-4o vs. o4-mini	119	14	98	145
grok-2 vs. o4-mini	112	16	105	143
grok-4 vs. o3	129	22	68	157
grok-4 vs. o4-mini	137	14	80	145
o3 vs. o4-mini	173	24	44	135

I. Cross-Model Overlap (Required, with images)

Table 22: Pairwise Jaccard overlap on majority-correct "solved" sets (Required items with images). Codes follow Table 2.

Model	*Non-reasoning*			*Reasoning*		
	gpt-4.1	gpt-4o	grok-2	grok-4	o3	o4-mini
Non-reasoning						
gpt-4.1	1.00	0.67	0.59	0.59	0.64	0.57
gpt-4o	0.67	1.00	0.63	0.57	0.55	0.52
grok-2	0.59	0.63	1.00	0.55	0.51	0.48
Reasoning						
grok-4	0.59	0.57	0.55	1.00	0.59	0.59
o3	0.64	0.55	0.51	0.59	1.00	0.72
o4-mini	0.57	0.52	0.48	0.59	0.72	1.00

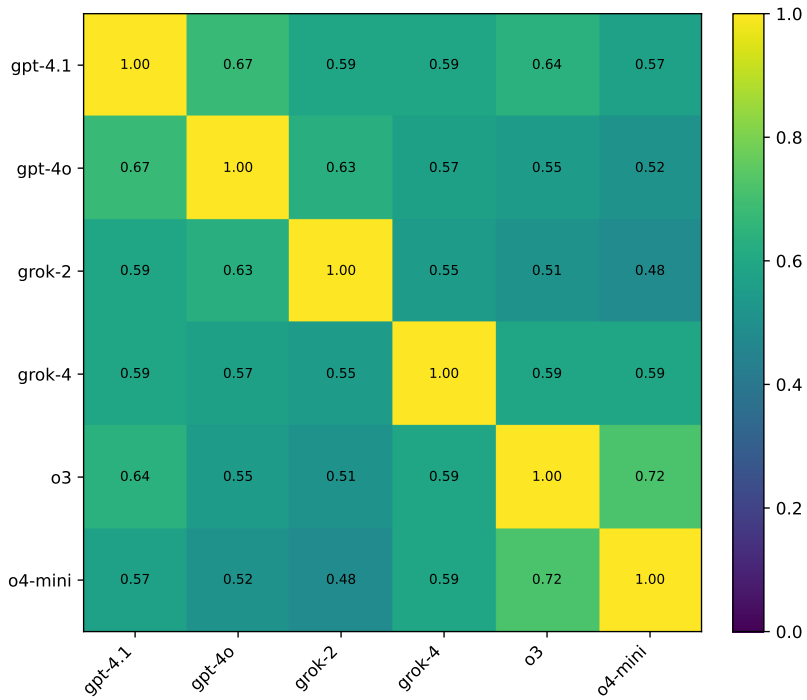


Figure 2: Jaccard overlap heatmap on *Required items with images* (majority-correct labels, $\geq 2/3$ runs).

J. Solvability without images (Required subset)

This appendix summarizes how often *Required* items were answered correctly when figures were not provided. For each item we aggregate across 18 runs (six models $\times$ three runs). An item contributes to Any ($\geq$ 1/3) if at least one run from at least one model is correct; Always (3/3) requires all three runs from a given model to be correct. The per-item source file is provided as `solved_counts/solved_count_combined_across_llms_without_images.csv`.

Table 23: Required, no image: items solved by exactly k of the six models under two criteria. Values are counts and percentages of the $N = 376$ Required items. Criteria: Any ($\geq$ 1/3 runs) and Always (3/3 runs). Δ is the difference in percentage points (Always $-$ Any).

k	**Any ($\geq$1/3)**		**Always (3/3)**		Δ **pp**
	Count	**Percent**	**Count**	**Percent**	
0	362	96.3%	366	97.3%	+1.1
1	3	0.8%	1	0.3%	-0.5
2	1	0.3%	5	1.3%	+1.1
3	6	1.6%	3	0.8%	-0.8
4	0	0.0%	0	0.0%	+0.0
5	2	0.5%	1	0.3%	-0.3
6	2	0.5%	0	0.0%	-0.5
Total	376	100.0%	376	100.0%	0.0

Note. Percentages may not sum to 100% due to rounding.

K. Uncertainty Details

Note. We also computed Wald intervals during development; they were essentially identical on these data, so we report bootstrap CIs only for brevity and consistency. See Table 3.
Family contrast (permutation) `agreement_out/family_contrast.txt` reports the observed effect (Reasoning $-$ Non-reasoning), the two-sided permutation p-value, and the null percentile band used for reference.

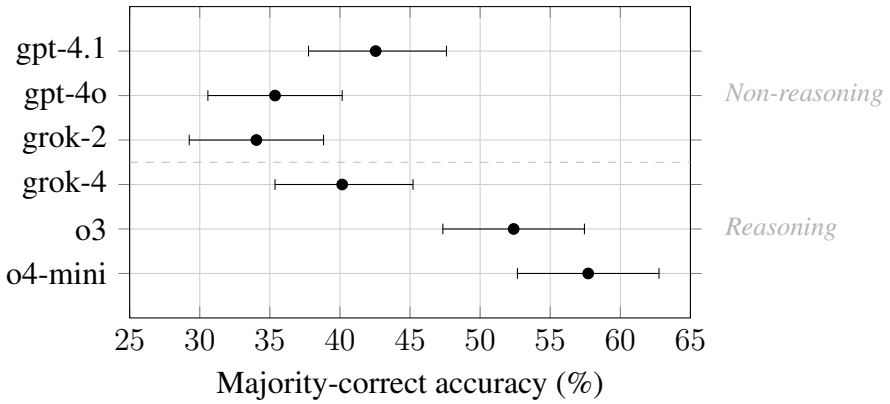


Figure 3: Majority-correct accuracy on *Required, with images* by model with 95% bootstrap CIs (10k resamples; items as unit).

K.1. Wilson confidence intervals for qualitative proportions

For qualitative proportions reported as x/n within an audit stratum (e.g., the fraction of responses labeled *Visual misread*), we compute two-sided 95% Wilson score binomial confidence intervals. Let $\hat{p} = x/n$ and $z = 1.96$ for 95% confidence. The Wilson interval is

$$\text{den} = 1 + \frac{z^2}{n}, \qquad \text{center} = \frac{\hat{p} + \frac{z^2}{2n}}{\text{den}}, \qquad \text{half} = \frac{z}{\text{den}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}.$$

We report $\text{center} \pm \text{half}$, converted to percentages. Wilson intervals behave well for small n and extreme counts (e.g., $x = 0$), yielding a nonzero upper bound when appropriate.

L. Original Qualitative Error Taxonomy (Before Merging)

Table 24 shows the original 7-category taxonomy used in initial coding. This includes a 7th category Diagram extraction, which was not in the first draft but was added during coding to distinguish failures in extracting visual symbols from misreading text.

All categories were later merged into the final 3-category taxonomy (Table 7) due to high rater disagreement and conceptual overlap.

Quick Reference Codes (used in coding):

`DATA` Data extraction `DIAG` Diagram extraction `REL` Diagram relations

`SCALE` Scale / axes `MULTI` Multi-representation `SYM` Symbolic manipulation

`EXPL` Explanation clarity

Table 24: Original qualitative error taxonomy (7 categories, used in coding).

Category	**Description**
Data extraction	Misreading explicit labels, numbers, or text embedded in the diagram.
Diagram extraction	Failing to extract and interpret visual symbols (e.g., dots, blocks, shaded units, objects on a scale, icons) that represent quantities or meanings.
Diagram relations	Misinterpreting geometric configurations (e.g., parallel/perpendicular, angle types, similar triangles, tick marks).
Scale / axes	Misreading graph scales, tick spacing, or unit mismatches (e.g., cm vs. m).
Multi-representation integration	Failing to combine information across subfigures or between diagram and table/graph.
Symbolic manipulation	Algebraic or procedural errors after correct visual extraction and integration.
Explanation clarity (pedagogy)	Correct answer with terse, non-pedagogical, or assumptive explanation that skips critical steps.