GRAPHTHEORY-VL: VISUAL INSTANCES FOR EXACT MATHEMATICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual mathematical problem solving requires reasoning with both diagrams and text. Yet models can answer some benchmark questions without their diagrams. For example, on MathVerse testmini, GPT-4V achieves 48.7% accuracy without diagrams, compared with 54.7% with diagrams and the same text (Zhang et al., 2025). This motivates checking what information a diagram contributes to an answer. We introduce GraphTheory-VL, a benchmark of 62 problems selected from 623 candidate question–image pairs through model screening and reference checks. This collection, Challenge62, includes a 20-problem subset, Hard20, selected for persistent screening failures. Each question links its diagram to an exact mathematical answer and an explanation of the instance facts needed to derive it. We evaluate three model families, generating four independent answers per problem. With the original images, mean accuracy across the four answers ranges from 10.89% to 31.85% on Challenge62 and from 1.25% to 5.00% on Hard20. We also evaluate the same models on Hard20 with the images removed and the text unchanged. Text-only mean accuracy ranges from 0.00% to 5.00%, with no consistent decline across models. Examining the missing instance facts and model responses distinguishes justified uncertainty from guesses that happen to match the answer. GraphTheory-VL thus supports evaluating mathematical difficulty together with whether a model's answer follows from the supplied visual information.

1 INTRODUCTION

Visual mathematical problem solving requires models to combine information from diagrams and text in mathematical reasoning. A solver may need to read vertex connections, count regions, or recognize a pattern before computing an answer. Benchmarks support progress on these tasks by comparing models and identifying which parts of a solution remain difficult.

Existing benchmarks span diverse visual contexts, competition problems, and university mathematics (Lu et al., 2024; Wang et al., 2024; 2026b). However, a problem containing an image does not necessarily test whether a model uses it. On MathVerse testmini (Zhang et al., 2025), the authors report 48.7% answer accuracy for GPT-4V without diagrams, compared with 54.7% when the same text is accompanied by the diagrams. On MathSight (Wang et al., 2026b), the authors likewise find that the same Qwen3-VL model performs better with the problem text alone than with the original image. These findings make visual dependence an empirical question: a benchmark should reveal what the image contributes and whether the model's solution respects that information.

Answer accuracy needs supporting evidence to resolve this question. Without an essential diagram, a model may reasonably decline to give a fixed answer, or guess a missing fact and happen to be correct. With the diagram present, it may compute correctly for the wrong object. Linking exact targets to visual instance facts and intermediate statements makes these different outcomes inspectable.

We introduce GraphTheory-VL to support this evaluation. The benchmark contains 62 visual mathematics problems, called *Challenge62*, including a nested *Hard20* subset selected for persistent screening failures. Selection uses a team-generated pool of 623 question–image pairs, archived model failures, and reference-quality checks. Targets span graph, algebraic, and topological computations. Each content card explains the relevant diagram facts and required answer, with reference

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

evidence for the supporting derivation or checks. Data, code, and archived responses are available for anonymous review.[1]

The benchmark remains challenging while allowing meaningful recovery across repeated attempts. We evaluate GPT-5.6 Luna, Gemini 3.6 Flash, and Claude Opus 4.8 with four separate responses per problem; Claude did not participate in screening. Their original-image mean accuracies range from 10.89% to 31.85% on Challenge62 and from 1.25% to 5.00% on Hard20. Across models and attempts, 43 of 62 problems are solved at least once, including only five hard problems. Reporting both at-least-one and all-attempt success distinguishes a recoverable failure from consistent correctness (Chen et al., 2021; Yao et al., 2025).

The image-removal study shows why low scores must be interpreted together with response behavior. On Hard20, paired accuracy differences are small and their intervals include zero. The reference audit instead identifies omitted instance facts, and analysis of all 120 first responses distinguishes justified uncertainty from unsupported answers. Among the 60 text-only first responses, 24 reasonably decline to determine the target. Across all 240 text-only scoring positions, the four target matches come from one model guessing the true value of an omitted parameter. These observations support a benchmark that makes visual dependence and answer support inspectable, rather than inferring them from a low text-only score.

GraphTheory-VL contributes the benchmark and linked references, 984 scored positions from a complete three-model comparison, and 120 behavioral annotations. Together, these assets support studying recurring failures, complementary model successes, and unsupported target matches in visual mathematical reasoning.

2 Related Work

GraphTheory-VL is most closely related to visual mathematical benchmarks, which cover questions presented with images, and graph reasoning benchmarks, which focus on computations over graph structures. We compare their inputs, targets, and evaluation methods.

Visual mathematical benchmarks. MathVista (Lu et al., 2024) contains questions in diverse visual contexts, while MATH-Vision (Wang et al., 2024) provides competition-level problems. To examine diagram use, the authors of MathVerse (Zhang et al., 2025) construct six versions of each problem with different amounts of textual and visual information. One pair presents identical text with or without the diagram, adding no replacement description. MathSight (Wang et al., 2026b) contains university-level mathematics problems. Its authors find that the same Qwen3-VL model performs better without the image than with the original image. These comparisons show that answering correctly does not by itself demonstrate diagram use. VCBench (Wang et al., 2025) provides problems with explicit visual dependencies. Other resources test robustness: DynaMath (Zou et al., 2025) provides generated problem variants, while the visual component of MathCheck (Zhou et al., 2025) covers changes to problem details, reasoning tasks, and visual styles. GraphTheory-VL links exact answers to the instance facts needed to derive them. Its reference explanations and recorded responses help distinguish uncertainty after image removal from unsupported target matches.

Graph reasoning benchmarks. GraphArena (Tang et al., 2025) provides graph computation tasks. VisionGraph (Li et al., 2024) and GVLQA, introduced with GITA (Wei et al., 2024), include graph images for evaluating structural understanding and reasoning. Visual Graph Arena (Babaiee et al., 2025) includes different layouts and isomorphism tasks, while VGCure (Zhu et al., 2025) targets fundamental visual graph understanding. Two closer comparisons concern mathematical content and answer checking. CombiGraph-Vis (Mahdavi et al., 2025) provides discrete olympiad problems with solutions and technique labels, and its authors evaluate repeated answers. GraphVerse (Sun et al., 2026) specifies eleven graph-image tasks, including shortest paths and coloring. Its authors generate answers algorithmically and score intermediate reasoning against graph structure. GraphTheory-VL includes algebraic and topological targets, such as group invariants and homology, beyond GraphVerse's stated task list. Our reference evidence connects each depicted instance to the computation used to check its complete answer.

[1] https://anonymous.4open.science/r/GraphTheory-VL-Experiments-2027/

Evaluating answers and their supporting reasoning. Beyond benchmark content, answer correctness and supporting reasoning require separate evaluation. The authors of We-Math (Qiao et al., 2025) use decomposed questions to examine component knowledge. Process supervision compares feedback on intermediate steps with final-answer feedback (Lightman et al., 2024), while description-based interventions examine perception in abstract reasoning (Wang et al., 2026a). We check explicit response statements against the supplied instance and reference evidence. This distinguishes target matches, justified uncertainty, and observable errors without attributing them to unobserved internal processes.

3 Benchmark Construction

3.1 From Candidate Creation to a Fixed Benchmark

GraphTheory-VL is built from team-generated mathematical questions and diagrams. Construction proceeds through candidate creation, model screening, reference verification, and quality filtering. Separate evaluation calls test the questions without the construction process's reference access, feedback, or computational checks.

Candidate creation. The team develops question–image pairs and stores each with an identifier, proposed target, and explanation. The source pool used for screening contains 623 pairs. The team creates and organizes the material with AI assistance. The retained records begin with the question workbook and preserve the item and later checks. They do not provide a complete per-question trace of the initial authoring interactions (Appendix A.1).

Screening for unresolved model performance. Archived screening contains four indexed positions for each of five model families. The Challenge rule selects questions on which both GPT-5.6 Luna and Gemini 3.6 Flash fail on their first response. The Hard rule selects questions with no observed correct response across all five models and four positions. DeepSeek V4 Flash Vision, Doubao Seed 2.1 Pro, and Qwen3-VL-32B-Instruct complete the screening panel. Challenge captures initial failures; Hard concentrates persistent observed difficulty.

Checking and correcting reference answers. Model-assisted review, AI-assistant inspection, derivation, and exact computation check whether the proposed target answers the depicted problem. When a reference changes, the existing screening answers are rejudged against it. Applying the screening rules to corrected labels gives 83 Challenge and 28 Hard candidates. Appendix A.3 records the corrections and their effects on fixed archived answers.

Quality filtering and retained membership. Sixteen unresolved question, image, or reference problems and five incomplete verifications are excluded. All 21 are Challenge candidates, and eight are also Hard candidates, leaving *Challenge62* with nested *Hard20*. Independent-evaluation success does not remove a retained problem. Quality decisions were not preregistered or documented as blinded, and some independent responses preceded them (Appendix A.10). Hard's screening condition means no observed success: one DeepSeek position for GT-0050 is missing, so it does not mean twenty completed failures for every problem.

3.2 What a Benchmark Item Contains

Each item pairs a question with a diagram and an accepted mathematical target. A content card states which visual facts matter, what the answer must contain, and where its reference evidence can be inspected. The collection covers graph, algebraic, and topological computations. Its 62 output forms comprise 19 integer tuples, 14 symbolic polynomials, ten symbolic tuples, eight integers, seven group-valued outputs, and four structured lists. For example, a tuple can require several counts, while a polynomial records more than one numerical property. Matching one component does not satisfy a request for the complete target.

Figure 1 illustrates how one omitted count affects an answer. The text describes a central polygon with an attached triangle per edge but leaves the sector count unspecified. The image fixes it at eight. Symmetries preserve this attachment pattern, so rotations and reflections refer to permutations of

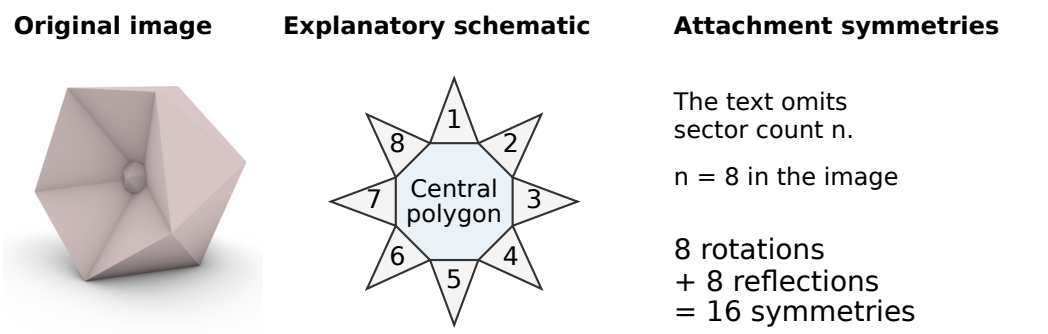

Figure 1: One visual fact can determine a requested answer component. The original GT-0514 image shows eight sectors around the central polygon. The text specifies the construction but omits this count. The labelled schematic explains the attachment pattern; it is an illustration, not the model input. Eight rotations and eight reflections preserve this pattern, giving 16 symmetries; six sectors would give 12. The full question also requests a cycle index and Euler characteristic, explained in Appendix B.2.

its sectors. A formula for an unspecified count does not determine the displayed instance's fixed answer.

3.3 Reference Evidence and Visual Dependence

Reference verification separates reading the instance from calculating its target. Exact arithmetic can check a graph count, matrix invariant, or polynomial after the relevant connections and labels have been recovered. It cannot by itself show that those connections match the drawing. The reference index therefore links structure claims to the question and image and calculations to their available evidence. Among the 62 retained items, 35 preserve supported targets and 27 use corrected targets. All have explanations and traceable evidence; the checks are model-assisted and computational, without dataset-wide independent human acceptance.

A separate text-sufficiency audit asks what is left unspecified when the image is removed. It identifies an omitted fact affecting every complete Hard20 target, while some components and general formulas remain derivable. The per-question arguments and their verification limits appear in Appendix H. Challenge62 also contains text-sufficient cases, so this Hard20 finding is not a claim of universal image necessity across the collection. GT-0560 retains a documented directional wording conflict, with a common-population exclusion analysis reported alongside the main results.

4 Evaluation Protocol

Models and inputs. We evaluate GPT-5.6 Luna, Gemini 3.6 Flash, and Claude Opus 4.8 with four separate responses per question and condition. Evaluation answers are separate from screening records; existing independent answers to identical inputs are reused. Each call solves the question without feedback from another attempt; four calls therefore measure repeated model responses, rather than an iterative agent with tools or human guidance. Requests provide the question and any options, but no reference, solution, prior response, or screening label. Solvers are instructed to avoid external tools and return `answer` and `reasoning` fields.

All models receive a 128,000-token output ceiling. GPT and Gemini request high reasoning effort, whereas Claude leaves effort unspecified. Sampling otherwise uses provider defaults. Realized computation and provider defaults can differ; Appendix A.4 gives the full configuration. The original-image condition O supplies the question and image. T removes the image while preserving

all wording, including the instruction referring to a supplied image. Its behavior is measured under that retained prompt.

Challenge62 is evaluated in O and Hard20 additionally in T, giving $62 \times 3 \times 4 + 20 \times 3 \times 4 = 984$ unique scoring positions. Hard O is included in Challenge O. There are 983 complete archived responses and one user-reported `Incorrect` decision for Claude T, GT-0560, response 2, whose inspected full answer is unavailable locally. Appendix A.6 reports the complete-archive sensitivity.

Answer evaluation. Correctness requires the complete requested target under the current reference, allowing established equivalence. Explicit proof or certificate requirements belong to that target. Otherwise, intermediate reasoning is assessed separately. T retains the original-target endpoint, so justified abstention can be appropriate behavior without a target hit. DeepSeek judgments, strict comparisons, and the user decision yield 181 `Correct` and 803 `Incorrect` labels. Additional target checks cover all 181 positive labels with no disagreement; they do not estimate false negatives or provide independent human calibration (Appendix A.5).

Metrics and uncertainty. For N questions, let $y_{ir} \in \{0, 1\}$ indicate correctness of indexed response $r \in \{1, 2, 3, 4\}$ to question i, and $c_i = \sum_{r=1}^{4} y_{ir}$. We report

$$\begin{aligned} \text{Acc@1} &= \frac{1}{N}\sum_i y_{i1}, & \text{MeanAcc} &= \frac{1}{4N}\sum_i c_i, \\ \text{Pass@4} &= \frac{1}{N}\sum_i \mathbf{1}[c_i > 0], & \text{Stable@4} &= \frac{1}{N}\sum_i \mathbf{1}[c_i = 4]. \end{aligned} \tag{1}$$

These established at-least-one and all-attempt criteria (Chen et al., 2021; Yao et al., 2025) use the same four responses without answer aggregation. Acc@1 fixes the first response. The distribution of c_i distinguishes occasional from repeated success.

MeanAcc intervals use 100,000 percentile bootstrap replicates over questions (Efron, 1979), keeping four responses together and resampling common questions jointly for comparisons. Binary question-level metrics receive Wilson intervals (Wilson, 1927). The bootstrap seed is 20260910. Comparisons are exploratory, unadjusted for multiplicity, and conditional on this selected benchmark. Appendix A.6 details all-zero groups and paired populations.

5 Independent Evaluation

The evaluation examines three properties of the retained benchmark: remaining difficulty, consistency across repeated attempts, and changes when the image is removed. All results in this section use independent answer records and fixed membership.

Current models leave room for improvement. The independent results show that the selected problems remain challenging for every evaluated model (Table 1). GPT achieves 79 correct responses out of 248, Gemini 71, and Claude 27. Their first-response accuracies are 21/62, 17/62, and 6/62, respectively. Mean accuracies are 31.85%, 28.63%, and 10.89%, respectively. The collection therefore retains difficulty without reducing every comparison to zero observed success. Independent successes also show that a screening failure need not recur.

Repeated attempts expand observed success, but they seldom produce success on every attempt. GPT solves 34 problems at least once and all four times on eight; Gemini reaches 27 and nine; Claude reaches 14 and one. Relative to each model's first independent response, the remaining three responses recover 13, 10, and eight additional problems. These counts measure recovery among the fixed four responses, without introducing an answer-selection procedure.

Hard20 retains a distinct difficulty profile. On Hard20, the models produce only four, two, and one correct original-image responses out of 80. GPT succeeds on three problems, Gemini on one, and Claude on one; no model answers any hard problem correctly on all four attempts. Appendix F gives the full success-count distribution.

The disjoint comparison with the other 42 Challenge problems shows a different profile. On that group, mean accuracy is 75/168 (44.64%) for GPT, 69/168 (41.07%) for Gemini, and 26/168

Table 1: Independent original-image evaluation. Rates and their intervals are percentages. Each model contributes four responses per problem. MeanAcc averages all four responses; Pass@4 and Stable@4 measure at least one and four correct responses. Intervals are 95% intervals computed at the problem level. Hard20 is contained in Challenge62.

Subset	Model	Correct	MeanAcc	Pass@4	Stable@4
Challenge62	GPT-5.6 Luna	79/248	31.85 [23.0, 41.1]	54.84 [42.5, 66.6]	12.90 [6.7, 23.4]
	Gemini 3.6 Flash	71/248	28.63 [19.4, 38.3]	43.55 [31.9, 55.9]	14.52 [7.8, 25.3]
	Claude Opus 4.8	27/248	10.89 [5.6, 16.9]	22.58 [14.0, 34.4]	1.61 [0.3, 8.6]
Hard20	GPT-5.6 Luna	4/80	5.00 [0.0, 11.2]	15.00 [5.2, 36.0]	0.00 [0.0, 16.1]
	Gemini 3.6 Flash	2/80	2.50 [0.0, 7.5]	5.00 [0.9, 23.6]	0.00 [0.0, 16.1]
	Claude Opus 4.8	1/80	1.25 [0.0, 3.8]	5.00 [0.9, 23.6]	0.00 [0.0, 16.1]

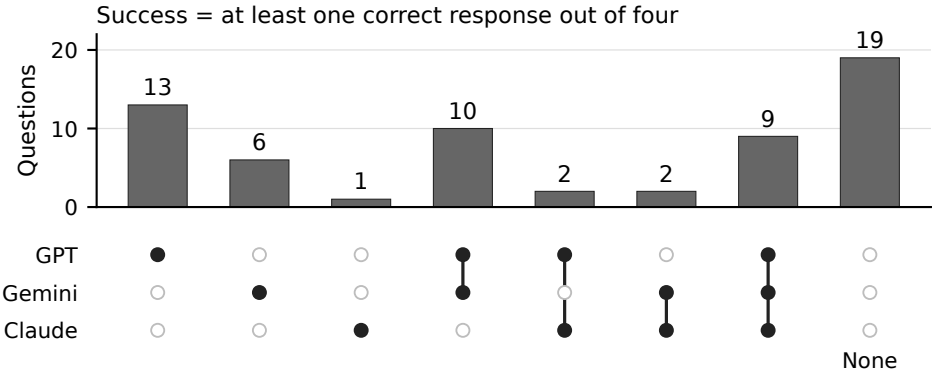

Figure 2: Model success sets on Challenge62. A filled dot marks a model with at least one correct response among its four attempts; each column is an exclusive combination. The seven nonempty combinations cover 43 problems, and the no-success column contains the other 19. These observed sets do not evaluate an ensemble or answer-selection rule.

(15.48%) for Claude. At least one model succeeds on 38 of these 42 problems, compared with five of the 20 hard problems. Thus, the broader collection offers opportunities to observe recovery, while the nested subset concentrates persistent failures. This contrast describes the sets produced by the screening procedure; it is not an estimate of the difficulty distribution of visual mathematics in general.

Success sets overlap only partly. Figure 2 shows that the models solve overlapping but different sets of questions. Across the three models and four responses, 43 of the 62 problems have at least one successful response. Nine problems are solved at least once by all three models, while 13 are solved only by GPT, six only by Gemini, and one only by Claude. Of the 19 problems with no observed success, 15 are in Hard20 and four are outside it. The union is an observational coverage statistic: selecting the correct response from this pool would require an additional decision rule that is not evaluated here.

The paired mean-accuracy difference between GPT and Gemini on Challenge62 is 3.23 percentage points, with a 95% interval of $[-6.85, 13.31]$. The GPT–Claude and Gemini–Claude differences are 20.97 points $[11.29, 30.65]$ and 17.74 points $[10.48, 25.81]$, respectively. These exploratory differences are conditional on the selected problems and service configurations.

Table 2: Descriptive Challenge62 original-image accuracy by existing output form. Model columns give MeanAcc in percent: correct responses divided by four times the Challenge question count. Hard gives the nested question count, not a separate scoring population. Categories contain four to nineteen questions; this post-hoc breakdown uses unchanged targets and labels.

Output form	Challenge	Hard	GPT	Gemini	Claude
Integer tuple	19	6	26.32	23.68	6.58
Symbolic polynomial	14	6	25.00	32.14	17.86
Symbolic tuple	10	3	47.50	42.50	10.00
Integer	8	2	31.25	28.12	15.62
Group-valued	7	2	42.86	25.00	10.71
Structured list	4	1	25.00	12.50	0.00
Total	62	20	31.85	28.63	10.89

Table 3: Hard20 image ablation. Text-only inputs remove the image while retaining the problem text and original instruction. Rates are percentages. Scores measure agreement with the original target. The final column compares the two conditions on the same 20 problems; all differences are in percentage points.

Model	O correct	T correct	O MeanAcc	T MeanAcc	Δ_{T-O} [95% CI]
GPT-5.6 Luna	4/80	0/80	5.00	0.00	-5.00 [-11.25, 0.00]
Gemini 3.6 Flash	2/80	0/80	2.50	0.00	-2.50 [-7.50, 0.00]
Claude Opus 4.8	1/80	4/80	1.25	5.00	+3.75 [-3.75, 15.00]

Difficulty spans several answer formats. The descriptive breakdown in Table 2 uses the six output categories already assigned in the content inventory. Integer outputs are single counts, integer tuples request multiple counts, and symbolic outputs require expressions; group-valued answers and structured lists retain additional mathematical structure. Even for the eight single-integer questions, the highest model mean is 31.25%. GPT's strongest category is symbolic tuples at 47.50%, while Gemini's mean on symbolic polynomials exceeds GPT's. This descriptive breakdown localizes errors without isolating output format as their cause.

Image removal cannot be read from the score alone. Both conditions produce few correct answers on Hard20. On Hard20, GPT and Gemini produce no correct text-only targets in 80 responses each. Claude produces four, all on one problem. The text-only minus original-image differences are -5.00, -2.50, and $+3.75$ percentage points. Their paired intervals all include zero (Table 3). More importantly, an incorrect text-only target may accompany a mathematically justified statement that the instance is unspecified. We therefore interpret this ablation together with the complete first-response analysis in Section 6.

Removing GT-0560, whose directional wording conflicts with its drawing, gives the same 19-question population for all models. Original-image MeanAcc becomes 5.26%, 2.63%, and 0%; Claude retains its four T hits on GT-0514. The pattern of sparse Hard success therefore persists. Appendix G gives all metrics and paired intervals; this exclusion does not replace evaluation with corrected wording.

6 What the Responses Reveal about Image Use

The image-removal results motivate examining what models do when the input does not determine a fixed answer. We analyze all 120 first responses on Hard20: 20 questions, three models, and two input conditions. Model-assisted review and AI-assistant adjudication check each response against the question, image, reference, and mathematical evidence. References and prior labels are visible at adjudication; Appendix A.7 details the protocol and uncertainty labels.

Table 4: Behavior of all first responses on Hard20. Each condition contains 60 responses, with 20 per model. O assigns one primary observable conflict or outcome per response; T assigns one final answering strategy. Attribution to the absent image is a separate dimension that can overlap the T strategies.

Category	GPT /20	Gemini /20	Claude /20	Total /60
O: observed primary category (one per response)				
Instance statement conflict	16	17	18	51
Modeling conflict	0	1	1	2
Condition conflict	0	0	1	1
Computation / logic conflict	0	1	0	1
Error stage unresolved	2	0	0	2
Target and checked reasoning consistent	2	1	0	3
T: final response strategy (one per response)				
Justified refusal	17	0	7	24
General formula / method	1	3	0	4
Final answer remains a guess	0	0	9	9
Missing image treated as empty	0	0	2	2
Unsupported instance answer	2	17	2	21
T: attribution to the absent image (separate dimension)				
Explicit attribution	1	5	2	8
Ambiguous attribution	1	4	1	6

6.1 Missing visual facts and reasonable uncertainty

The text-sufficiency audit identifies an omitted instance fact for each of the 20 complete Hard targets. Its test is whether changing a fact left unspecified by the text can change the requested answer. For example, changing a diagram's sector count changes its number of symmetries, and changing a graph's edge connections can change a requested count. Appendix H lists the fact and alternative-instance argument for every question, distinguishing explicit alternatives from arguments that still lack a fully specified construction.

An undetermined complete answer can still have computable components. In the example in Figure 1, the text describes a polygon with one attached triangle per edge, but does not give the number of sectors. It supports a formula in terms of that number, while the image fixes a numerical instance. Recognizing this distinction is a useful response to image removal. The finding concerns Hard20; Challenge62 also includes text-sufficient questions.

6.2 Text-only errors reflect different answering strategies

Table 4 shows that target errors have different meanings under missing input. Of 60 text-only responses, 24 reasonably decline to determine the target and four give a general formula or method. These 28 responses retain relevant uncertainty even though they do not match the fixed answer. Nine explicitly retain a guess, 21 commit to instance facts unsupported by the text, and two treat the missing image as an empty mathematical object. The latter interpretation introduces an object that the prompt never specified.

The strategies vary across the sampled first responses: GPT contributes 17 justified refusals, Gemini contributes 17 unsupported instance answers, and Claude contributes all nine final guesses. Aggregate accuracy does not distinguish productive uncertainty from unsupported commitment.

Some responses also claim to have seen evidence that was not supplied. Eight of 60 explicitly attribute a fact to the absent image, with six additional cases left ambiguous. This is different from making an unsupported numerical assumption: the annotation requires an explicit claim of visual evidence. The prompt retains its reference to a supplied image, so these frequencies are conditional on that wording.

6.3 One missing count explains every text-only target match

The sector example makes the limitation of target-only scoring concrete. Its attachment pattern has 16 symmetries when there are eight sectors, compared with 12 when there are six. The full target also asks how those symmetries permute the sectors and for a topological count. It therefore needs the particular value eight, even though some components can be derived symbolically.

Claude's first text-only response guesses eight and later states, "The rendering shows n = 8 sectors." Its conditional calculation matches the complete target, but its visual claim cannot follow from the input. All four Claude text-only successes occur on this question. GPT's first response keeps the sector count symbolic and misses the numerical target while preserving the missing information. The scoring rule treats these as a target hit and a miss, respectively; the behavior analysis explains why the miss can still be an appropriate mathematical response. Appendix B.2 gives the full question structure, target, and calculation.

6.4 Providing the image does not ensure a supported solution

Original-image responses frequently make claims that conflict with the depicted instance. Among the 60 first responses, 51 contain an explicit instance-statement conflict. Two contain modeling conflicts, one violates a mathematical condition, and one makes a computational error after identifying the relevant graph. Two errors cannot be localized, while three responses match the target and the derivation examined. These are conflicts in written answers, rather than direct measurements of internal perception.

A conflicting instruction affects part of this analysis. GT-0560 contains directional wording inconsistent with its drawing; removing its three original-image responses leaves 49 instance conflicts among 57 responses. The uniform exclusion analysis in Appendix G keeps this input defect separate from model error. Exact reasoning can fail even after the relevant structure is identified: a response in GT-0411 recognizes the graph but returns the wrong algebraic structure, despite matching its size. Appendix B explains this example and three related cases in full, including why a partial numerical check can accept an incomplete target.

These observations connect target scores to the facts used to justify them. They distinguish missing information, unsupported instance assumptions, and mathematical errors that require different improvements.

7 Discussion and Conclusion

GraphTheory-VL provides a benchmark for studying whether mathematical answers follow from the visual instances supplied to a model. Its 62 problems combine exact targets with reference evidence and complete available responses. Independent evaluation shows both recovery across attempts and persistent difficulty, particularly on the nested Hard20 subset. The response analysis explains why a correct final answer can coexist with an unsupported assumption, and why a justified refusal can miss the benchmark's fixed target.

The resource supports evaluating improvements at identifiable stages of a solution. Solvers can expose inferred connections or counts before their calculations are judged, and retain an unresolved parameter when the input omits it. The content cards and observed responses provide cases for these investigations.

The conclusions are conditional on model-based selection, a small hard subset, and the recorded third-party service configurations. Original-image and text-only scores are both low on Hard20, and their paired differences do not establish a uniform image-removal effect. Reference and behavior checks use AI assistance and exact computation without independent human calibration; positive-label checks do not estimate missed correct answers. Reference corrections affect fixed-response scores, and some independent responses preceded final quality decisions. Structural-input interventions and later-attempt behavior remain unmeasured. Within this scope, GraphTheory-VL offers an inspectable link between benchmark difficulty, the information supplied by a diagram, and the evidence behind a model's answer.

AI USE STATEMENT

Generative AI tools were used in developing the research framing and experimental design, implementing evaluation and analysis code, checking and revising mathematical references, formulating and checking mathematical arguments, organizing the data, analyzing responses, interpreting results, and drafting this manuscript. DeepSeek supplied semantic answer judgments; Claude supplied preliminary qualitative assessments; Codex assisted with data integration, exact-computation checks, evidence examination, analysis, and writing. Human review is recorded only where it was explicitly supplied. Reference and response checks combine source-linked inspection with exact computation, as described in Sections 3 and 4 and Appendix A. These procedures are not represented as independent human verification of the full collection.

REPRODUCIBILITY STATEMENT

The evaluation is specified by fixed problem memberships, reference targets, input conditions, model identifiers, and four indexed responses per problem. Appendix A describes the scoring and statistical procedures, and Appendices C and D give per-problem outcomes and the content index. Data, experimental code, and recorded responses are accessible through the anonymous repository linked in Section 1. The artifact includes the current question–image pairs, full available responses, label sources, reference evidence, and the earlier screening and evaluation records supporting selection and correction. Its companion index supports per-question inspection (Appendix A.11). Credential-free code, exact prompts, and configurations support rerunning the procedures with separately supplied service access. The archive preserves the distinction between 983 complete current responses and one user-adjudicated position, and between checker source and recorded execution. The retained benchmark questions and diagrams were created by the team.

REFERENCES

Zahra Babaiee, Peyman Kiasari, Daniela Rus, and Radu Grosu. Visual graph arena: Evaluating visual conceptualization of vision and multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 2081–2113. PMLR, 2025. URL https://proceedings.mlr.press/v267/babaiee25a.html.

N. L. Biggs. Chip-firing and the critical group of a graph. *Journal of Algebraic Combinatorics*, 9 (1):25–45, 1999. doi: 10.1023/A:1018611014097. URL https://link.springer.com/article/10.1023/A:1018611014097.

Seth Chaiken. A combinatorial proof of the all minors matrix tree theorem. *SIAM Journal on Algebraic and Discrete Methods*, 3(3):319–329, 1982. doi: 10.1137/0603033. URL https://epubs.siam.org/doi/10.1137/0603033.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Yunxin Li, Baotian Hu, Haoyuan Shi, Wei Wang, Longyue Wang, and Min Zhang. VisionGraph: Leveraging large multimodal models for graph theory problems in visual context. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27903–27919. PMLR, 2024. URL https://proceedings.mlr.press/v235/li24ab.html.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/663bce02a0050c4a11f1eb8a7f1429d3-Abstract-Conference.html.

Hamed Mahdavi, Pouria Mahdavinia, Alireza Farhadi, Pegah Mohammadipour, Samira Malek, Majid Daliri, Pedram Mohammadipour, Alireza Hashemi, Amir Khasahmadi, and Vasant Honavar. CombiGraph-Vis: A curated multimodal olympiad benchmark for discrete mathematical reasoning. In *MATH-AI: The 5th Workshop on Mathematical Reasoning and AI at NeurIPS*, 2025. URL https://neurips.cc/virtual/2025/131041.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, YiFan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. We-Math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20023–20070. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.983. URL https://aclanthology.org/2025.acl-long.983/.

Richard P. Stanley. Smith normal form in combinatorics. *Journal of Combinatorial Theory, Series A*, 144:476–495, 2016. doi: 10.1016/j.jcta.2016.06.013. URL https://www.sciencedirect.com/science/article/pii/S0097316516300474.

Yuanfu Sun, Yuanhang Ren, Kang Li, Chuanhao Ji, Jiaxi Li, Jiajin Liu, Ninghao Liu, and Qiaoyu Tan. GraphVerse: A comprehensive visual graph reasoning benchmark for multimodal large language models, 2026. URL https://arxiv.org/abs/2608.06769. Version 1.

Jianheng Tang, Qifan Zhang, Yuhan Li, Nuo Chen, and Jia Li. GraphArena: Evaluating and exploring large language models on graph computation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/77fa8253adfc8b33209639f3e9985741-Abstract-Conference.html.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95095–95169. Curran Associates, Inc., 2024. doi: 10.52202/079017-3014. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad0edc7d5fa1a783f063646968b7315b-Paper-Datasets_and_Benchmarks_Track.pdf.

Xinhe Wang, Jin Huang, Xingjian Zhang, Tianhao Wang, and Jiaqi W. Ma. Your reasoning benchmark may not test reasoning: Revealing perception bottleneck in abstract reasoning benchmarks. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18113–18124, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.826. URL https://aclanthology.org/2026.acl-long.826/.

Yuandong Wang, Yao Cui, Yuxin Zhao, Zhen Yang, Yangfu Zhu, and Zhenzhou Shao. MathSight: A benchmark exploring have vision-language models really seen in university-level mathematical reasoning? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47573–47604. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.acl-long.2198. URL https://aclanthology.org/2026.acl-long.2198/.

Zhikai Wang, Jiashuo Sun, Wenqi Zhang, Zhiqiang Hu, Xin Li, Fan Wang, and Deli Zhao. Benchmarking multimodal mathematical reasoning with explicit visual dependency, 2025. URL https://arxiv.org/abs/2504.18589.

Yanbin Wei, Shuai Fu, Weisen Jiang, Zejian Zhang, Zhixiong Zeng, Qi Wu, James T. Kwok, and Yu Zhang. GITA: Graph to visual and textual integration for vision-language graph reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 44–72. Curran Associates, Inc., 2024. doi: 10.52202/079017-0002. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/00295cede6e1600d344b5cd6d9fd4640-Paper-Conference.pdf.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1927.10502953.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, pp. 9965–10017, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1b126cc38b8638e07bef37e7b2bb72bf-Paper-Conference.pdf.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MATHVERSE: Does your multi-modal LLM truly see the diagrams in visual math problems? In *Computer Vision – ECCV 2024*, pp. 169–186, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73242-3. doi: 10.1007/978-3-031-73242-3_10. URL https://link.springer.com/chapter/10.1007/978-3-031-73242-3_10.

Zihao Zhou, Shudong Liu, Maizhen Ning, Wei Liu, Jindong Wang, Derek Wong, Xiaowei Huang, Qiufeng Wang, and Kaizhu Huang. Is your model really a good math reasoner? evaluating mathematical reasoning with checklist. In *International Conference on Learning Representations*, pp. 34238–34281, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/54d2d38a56a74387d5916ee40e462295-Paper-Conference.pdf.

Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. Benchmarking and improving large vision-language models for fundamental visual graph understanding and reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30678–30701. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1482. URL https://aclanthology.org/2025.acl-long.1482/.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. DynaMath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *International Conference on Learning Representations*, pp. 48337–48383, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/78b248ea6f627431bba5029d92be8a3d-Paper-Conference.pdf.

A Additional Construction and Evaluation Details

A.1 Team Authoring and the Available Construction Record

The questions are generated by the team, rather than selected as existing questions from external papers. The benchmark therefore does not use a source-paper count or journal-quartile inclusion rule. The initial graph-question workbook contains 628 items; five recorded input or reference exceptions leave the 623-item screening pool. The internal lineage links every retained question and image to a unique source-workbook row. This is item-level source-workbook lineage, distinct from an authoring transcript or the reference checks performed later.

The available record begins with drafted questions, diagrams, proposed answers, and explanations. It does not preserve a complete initial authoring transcript, per-item authoring model configuration, or iteration budget. The reproducible portion covers the fixed source pool, archived screening, reference corrections, quality decisions, independent responses, and analysis. Construction permits human direction, AI assistance, and mathematical checks; evaluation instead gives each solver the fixed prompt and input without the reference or feedback from previous attempts. Repeating that isolated call four times does not evaluate the full construction workflow as an agent. The present work does not establish a measured rate of generating additional valid problems.

A.2 Screening, Quality Decisions, and Fixed Membership

The source pool contains 623 identified questions, their images, reference answers, and archived screening records. The five requested screening identifiers are `deepseek-v4-flash-vision-exp`, `doubao-seed-2-1-pro-260628`, `gemini-3.6-flash`, `gpt-5.6-luna`, and `qwen3-vl-32b-instruct`. Four scheduled positions per model produce 12,460 positions. With the active reference corrections, these contain 5,107 correct, 7,345 incorrect, and eight missing labels. Screening records and independent evaluation records are maintained separately.

Before quality exclusions, the two first-response failures identify 83 Challenge candidates and the zero-observed-success condition identifies 28 Hard candidates. The 16 questions with unresolved problems and five awaiting reference verification are excluded from evaluation membership. All 21 belong to Challenge, and eight also belong to Hard. The retained Hard20 screening records contain 399 incorrect positions and one missing position, GT-0050's third DeepSeek response, which remains missing.

The five references awaiting verification are GT-0322, GT-0345, GT-0347, GT-0349, and GT-0598. Their exclusion concerns incomplete evidence rather than a confirmed wrong answer. Other exclusions have question–image conflicts or refuted targets without resolved replacements. Membership follows these audit decisions and remains fixed when reporting independent evaluation responses.

A.3 Reference Corrections and Mathematical Evidence

The fixed 134-question audit contains 78 reference corrections, 35 supported references, 16 unresolved problems, and five incomplete verifications. Two additional corrections, GT-0148 and GT-0625, were recorded separately, giving 80 in total. The audit combines model responses with direct image examination, mathematical derivation, and executable checks by an AI assistant. Among retained Challenge questions, 35 original targets are supported and 27 have been corrected. Every retained question has an explanation and a location for the mathematical evidence used to support its current target.

Evidence takes a form appropriate to the object. Exact arithmetic checks matrix invariants, integer products, and coefficients of identical polynomial monomials. Invariant-factor questions require group structure, not merely group order. Image interpretation remains separate: exact computation on an incorrect adjacency matrix would not validate the source instance. The reference index links structure claims to images or question rules and computations to their outputs.

The 78 corrections prompted 27 rejudgment batches covering 2,671 complete archived answers: 1,559 screening and 1,112 independent answers. Strict offline comparisons resolved 1,233, and DeepSeek evaluated the other 1,438. The labels contain 633 correct and 2,038 incorrect answers,

with 775 changes, including 255 independent-evaluation changes. The two earlier corrections add 40 screening rejudgments. These purposively selected records do not estimate the error rate of the full source collection.

Reference sensitivity uses fixed question memberships and identical answer texts before and after correction and rejudgment. Neither new answers nor changes in membership enter that comparison. Its interpretation includes both reference changes and the application of a new judgment to an existing response. It is consequently separate from the independent Challenge62 and Hard20 results.

A.4 Generation Prompts, Services, and Technical Recovery

The exact common system message is:

> You are solving an expert-level graph theory problem. Use both the problem text and the supplied image. Work independently and do not use external tools or external information. Return the final answer in the required format.

The user message begins with `Problem:`, a newline, and the complete question. Nonempty options are appended after a blank line as `Options:`, a newline, and their contents. After a blank line, every message ends with:

```
Return your response using exactly these two fields:
answer:  <final answer>
reasoning:  <concise explanation>
```

The companion contains the exact expanded messages for all 62 questions, verified against the stored prompt identifiers of all 983 complete responses. T removes the image block without changing either textual message. Statements about an absent image are therefore observed under an instruction that still refers to a supplied image.

The independent identifiers are `gpt-5.6-luna`, `gemini-3.6-flash`, and `claude-opus-4-8`. GPT uses `max_output_tokens=128000`, while Gemini and Claude use `max_tokens=128000`. GPT and Gemini request high reasoning effort, whereas Claude leaves effort unspecified. Temperature and nucleus-sampling settings use provider defaults. Records retain requests, responses, finish status, and available usage information. The task-order seed 20260823 is not transmitted as an API sampling seed.

Krill serves GPT and Claude requests, BUZZ serves the initial Gemini requests, and OpenLux serves Gemini recovery and further coverage as well as GPT recovery. The final 280-answer completion of the common Challenge table comprises 19 GPT responses through Krill, 121 GPT responses through OpenLux, and 140 Gemini responses through OpenLux. The same requested model, prompt, image, and output ceiling are retained when changing the service for these positions. Service changes are recorded alongside model identifiers because the providers' underlying routing and weights are not independently verified.

Technical failures and empty responses are recovered as missing positions, retaining failed attempts separately. A completed response is not replaced because it is incorrect. Challenge O contains 744 complete answers, comprising 464 reused independent answers and 280 additional answers, while Hard T contributes 239 complete answers and one manual label. Their 984 unique scoring positions count Hard O once, within Challenge O.

A.5 Judgment Semantics and Additional Checks

DeepSeek V4 Flash is the semantic judge, accessed under `deepseek-v4-flash`. The corrected-reference and completion prompts supply the current question, target and explanation, and full candidate response, without the previous correctness label. They request a target verdict, a separate reasoning status, and a mathematical rationale. For GT-0019, the target requires a valid irreducibility argument. For a question requesting a polynomial alone, unrelated intermediate errors are recorded separately from target equivalence.

Strict comparisons require an explicit complete target in a supported form, including ordered integer tuples and restricted exact integer expressions. Unclear extractions or unestablished equivalences

are not automatically marked correct. The registry binds the 983 automatically scored positions to their question, current reference, full response, and judgment source; the manual exception binds to the user's decision record. Methods total 832 semantic judgments, 151 strict or deterministic comparisons, and one manual decision, with 344 positions covered by corrected-reference rejudgment.

Additional checks cover all 181 current correct responses, with no final-target disagreement. They compare explicit answers with visible references, using exact calculations where applicable, without independently validating every intermediate statement. This positive-only coverage cannot estimate the false-negative rate among 803 incorrect labels or exclude shared errors in references and judgments. A blinded, independently adjudicated sample of both labels has not been collected. An earlier audit of 396 responses received 394 model reviews and raised 33 issues. Thirty-two were adjudicated and one belongs to an excluded question. The audit overlaps 147 current positions, of which 62 retain the same reference. This is coverage information, not human calibration of judge accuracy.

A.6 Statistical Units and Question Exclusion

Each question contributes four binary outcomes per model and condition. Bootstrap samples draw questions with replacement, carry their four outcomes together, and use the same indices on both sides of a comparison. The 95% percentile intervals use 100,000 replicates and linear interpolation of quantiles. Binary question-level metrics use Wilson intervals. Pass@4 and Stable@4 measure observed any-attempt and all-attempt success without selecting an answer for deployment.

All-zero groups produce a degenerate percentile interval for MeanAcc because every empirical resample remains zero. This does not establish a zero underlying success probability. With 20 questions and no observed success, the Wilson upper endpoint for Pass@4 is 16.11%. Model and O/T differences are exploratory and have no multiplicity correction. All intervals describe variation across this selected set rather than uncertainty from a random sample of all visual mathematics.

The principal paired Hard comparisons include the user-reported label described below. Appendix G removes GT-0560 from every model and condition, addressing both its incomplete archive and its conflicting wording on a common population. Excluding the whole question preserves four responses per condition; no answer or correctness label is changed.

A.7 Full-Response Behavioral Coding

Behavioral analysis fixes response 1 for every Hard question, model, and condition before inspecting behavior, yielding 120 complete answers. Twenty question packs contain the image, text, current reference, mathematical evidence, and six responses. The Claude initial review sees randomized candidate letters and O/T conditions with model identity and original judgment hidden. Twenty reviews complete with a 16,000-token ceiling. Final AI-assistant adjudication sees the complete answers, references, original labels, and review suggestions, and is not blinded.

Text sufficiency is assessed per question by identifying an unspecified instance fact or compatible instances with different targets. A reference to a figure alone is insufficient evidence of underdetermination. Response coding distinguishes target hits, the stance of the final answer, specific conflicts with instance or mathematical facts, and explicit attribution of facts to an unprovided image. The summary assigns one primary O conflict and one primary T strategy per answer, while visual attribution is a separate dimension. Each positive claim is supported by a located quotation and corresponding source or mathematical evidence. The 134 indexed quotations have been matched to their full responses. Claude's involvement in both solving and initial review, and the reference exposure of the final assistant checks, make these model-assisted annotations rather than independent human judgments. Quotation agreement establishes traceability, not the accuracy of a category assignment. No blinded independent recoding or inter-annotator agreement estimate is available; the category frequencies remain conditional on this protocol.

A.8 GT-0560 Wording and Manual Judgment

GT-0560 describes two curved arrows as issuing from vertex 4. The drawing instead contains $4 \to 2$ and $3 \to 4$, whose entries follow the explicit tail-to-arrowhead convention. The reference uses these

displayed directions. The evaluated wording is preserved and the inaccurate phrase documented as an erratum. This is a defect in the supplied item: a response compatible with the phrase cannot be treated as an unambiguous image-reading error. None of the three O first responses explicitly says that this phrase determined its chosen direction. Gemini's relevant matrix entries are compatible with it, but the causal basis of those entries cannot be inferred from the answer. The uniform exclusion in Appendix G therefore complements the main scores without assigning the discrepancy solely to the model.

Claude's second T response for this question has three incomplete local stream records. The user inspected an answer and explicitly judged it incorrect, but that complete inspected answer was not supplied to the local archive. We retain the decision as one user-reported manual label, without counting it as a complete response or a DeepSeek judgment. This is separate from Claude's second O response on the same question, which is correct. The archive distinction is retained in both the principal scores and the complete-archive sensitivity above.

A.9 Scope of the Closest-Resource Comparison

Our comparison examines task definitions and public assets, not a cross-benchmark ranking. MathVerse's inspected testmini archive contains 788 distinct problems; 787 text-only prompts exactly match the corresponding Text Dominant prompt. Image removal in that condition generally does not replace the omitted diagram with new text. MathSight supplies original, hand-drawn, photographed, and image-free conditions. Its inspected 661-question file includes algebra, topology, and discrete-mathematics examples as well as its larger calculus component. CombiGraph-Vis examples cover discrete olympiad techniques and include image-bearing jump-counting, coverage, coloring, and tiling problems. GraphVerse's eleven specified tasks have algorithmic references and process-sensitive evaluation. These comparisons acknowledge overlapping evaluation dimensions and identify the requested mathematical targets and reference assets as the relevant distinctions (Zhang et al., 2025; Wang et al., 2026b; Mahdavi et al., 2025; Sun et al., 2026).

A.10 Recorded Chronology and Review Visibility

The records separate screening responses from independent evaluation responses, but final quality decisions follow some independent calls. Timestamped screening logs span August 23–September 3, 2026; no corresponding per-request Doubao log was located. Of the 983 complete responses in the current registry, 564 were recorded on September 8–9 and 139 during Hard extension on September 10, 02:21–04:47 UTC. These 703 responses preceded application of the final membership. Reference review, correction, old-answer rejudgment, and the Hard extension overlapped on September 10. The quality-exclusion policy was recorded at 04:47 UTC and membership applied at 04:58 UTC. The remaining 280 Challenge responses were recorded later that day, 15:53–17:47 UTC. The registry assembly date, September 11, is distinct from these response dates.

All 134 reference-review API packets supplied the image and reference but no candidate answer or old-label field. The local correction workflow had broader evidence: independent response texts, model identities, prior judgments, and rejudgment comparisons. The release policy states that independent performance is not an input to the screening rule or quality gate. However, there is no complete access log for the final decision maker, and the gate was not preregistered. These records support separation of answer sets and a documented membership rule; they do not establish global blinding or prospective fixation before independent evaluation. The companion's timeline preserves the exact timestamps and packet-field audit.

A.11 Anonymous Artifact and Audit Entry Points

The anonymous repository linked in Section 1 provides the experimental code, records, and a portable audit companion. Its README identifies the current results and gives the archive extraction and workspace restoration commands. After extraction, the companion opens at `companion/index.html`. Each question page connects the original text and image, current target, content card, reference evidence, four indexed responses per evaluated model and condition, and judgment sources. `questions/index.json` indexes all 62 entries, while `responses/slots.jsonl`

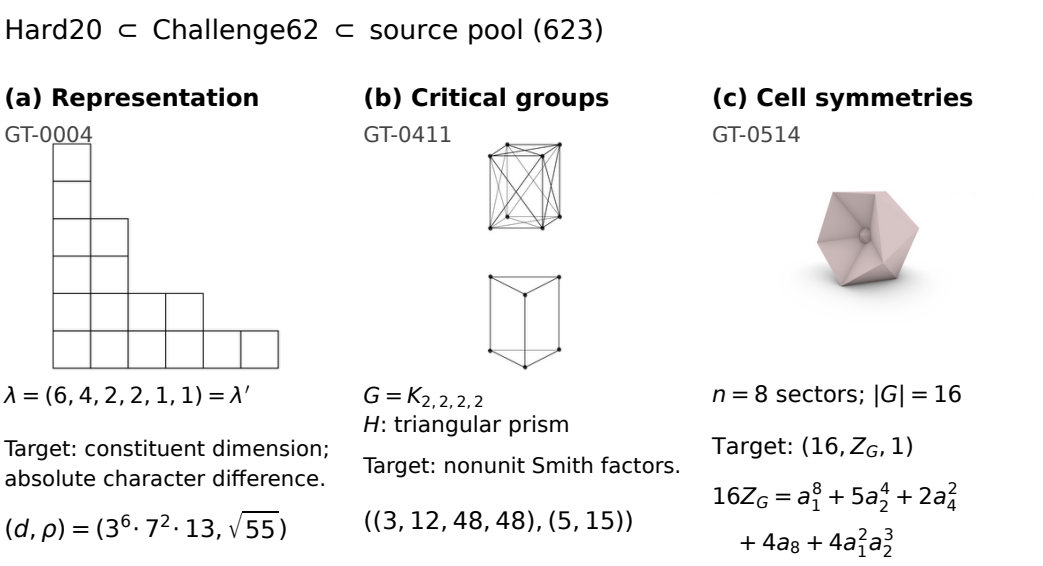

Figure 3: Additional resource examples. Each panel is a separate question, pairing its original image with its requested mathematical target. These examples illustrate different outputs; they are not three components of one item. Challenge62 includes the nested Hard20 subset.

indexes 984 distinct positions. The answer directory contains 983 complete response texts. The remaining position contains the explicit manual-label record and no fabricated answer.

`protocol/prompts/` stores all exact expanded solver prompts; `protocol/configurations/` stores the 17 located configurations for these answers. The observed UTC response times and source bindings are retained. `provenance/source_inventory.json` lists copied evidence, including the 121 directly linked reference assets, and `provenance/known_gaps.json` records closure limits. The revision analyses add uniform item exclusion, per-question text-sufficiency arguments, the 12-response case census, and the chronology audit. Paths in the companion resolve within its directory; credentials and machine-specific author paths are omitted.

The companion's offline verifier checks slot totals, full-answer coverage, links, and packaged-file integrity. It does not rerun every historical mathematical checker or certify every visual reconstruction. Some multi-question checker scripts depend on inputs outside the current 62-question set; source availability is distinguished from a self-contained execution receipt. The repository also preserves earlier screening and independent-response records needed to trace membership and rejudgment. Large records are distributed in archive parts with a restoration script; credentials and author paths are removed from the release copy while scientific values and labels are retained. Exact reproduction of service calls requires the reader's own credentials and available provider endpoints. Original creator attribution and redistribution rights are not established by anonymous hosting.

B Exact targets and illustrative response checks

The main text uses one sector-count example to explain the relation between a diagram and a fixed answer. This appendix gives the full calculation and the additional algebraic and graph examples. Figure 3 displays their original images with item-specific descriptions; Figure 4 contrasts an unsupported instance assumption with an incorrect computation after instance identification.

B.1 Equal group order does not determine the critical group

The critical group is an algebraic structure associated with a graph, expressible as a list of integer factors. Its size equals the number of spanning trees, but different factor lists can describe different

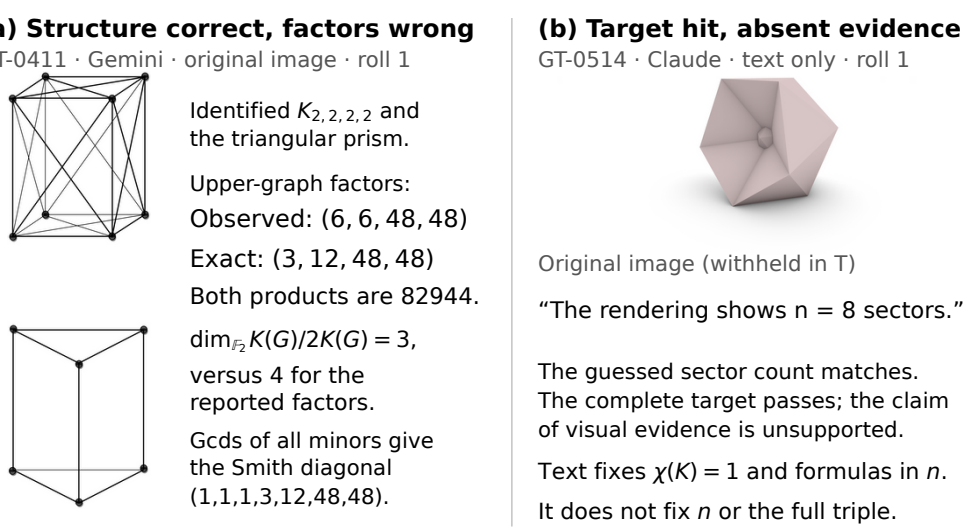

Figure 4: Detailed examples of target–evidence disagreement. In GT-0411, the correct graph identity precedes incorrect group factors. In GT-0514, a text-only answer guesses the correct sector count and attributes it to a rendering that was absent from its input. The subsections below define the mathematical quantities and explain the checks.

groups of the same size. The graph Laplacian used here is a finite matrix: diagonal entries are vertex degrees and off-diagonal entries are negative edge multiplicities. Removing one row and the corresponding column gives a reduced Laplacian. Smith normal form diagonalizes this integer matrix using invertible integer row and column operations; its nonunit diagonal entries give the factors.

For a connected graph with reduced Laplacian $\widetilde{L} \in \mathbb{Z}^{(n-1)\times(n-1)}$, the critical group can be represented as

$$K(G) = \mathbb{Z}^{n-1}/\widetilde{L}\mathbb{Z}^{n-1}.$$

If the Smith diagonal is $d_1, \ldots, d_{n-1}$, then $K(G)$ is the direct sum of the cyclic groups $\mathbb{Z}/d_i\mathbb{Z}$, omitting the factors $d_i = 1$. The product of the factors equals $\det \widetilde{L}$, the spanning-tree count (Biggs, 1999; Stanley, 2016; Chaiken, 1982).

For the upper graph in GT-0411, the accepted structure is $K_{2,2,2,2}$: eight vertices partitioned into four pairs, with an edge between every two vertices in different pairs and no edge within a pair. Its reduced Laplacian is

$$\widetilde{L} = \begin{pmatrix} 6 & 0 & -1 & -1 & -1 & -1 & -1 \\ 0 & 6 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 6 & 0 & -1 & -1 & -1 \\ -1 & -1 & 0 & 6 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & 6 & 0 & -1 \\ -1 & -1 & -1 & -1 & 0 & 6 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & 6 \end{pmatrix}.$$

Let Δ_k be the gcd of its $k \times k$ minors, with $\Delta_0 = 1$. Direct integer computation gives

$$(\Delta_0, \ldots, \Delta_7) = (1, 1, 1, 1, 3, 36, 1728, 82944).$$

Taking successive quotients produces the Smith diagonal $(1, 1, 1, 3, 12, 48, 48)$. The nontrivial factors are therefore $(3, 12, 48, 48)$. The candidate factors $(6, 6, 48, 48)$ have the same product, but $K/2K$ would have dimension four over $\mathbb{F}_2$, whereas the accepted group has dimension three. This gives a short independently checkable reason why group order alone cannot validate the candidate. The lower graph is a triangular prism, with accepted nontrivial factors $(5, 15)$.

B.2 AN UNSPECIFIED SECTOR COUNT AND A CORRECT CONDITIONAL CALCULATION

A symmetry maps the arrangement to itself while preserving the central face. For this item, the requested cycle index records how each symmetry permutes the surrounding sectors: a term such as a_2^4 describes four cycles of length two. Averaging these terms over all symmetries gives the cycle-index polynomial. The Euler characteristic is the vertex count minus the edge count plus the face count. These are the two additional quantities beyond the symmetry count illustrated in the main text.

The text of GT-0514 defines a central n-gon with one attached triangle per boundary edge and a distinguished central face. Its cell counts are $V = 2n$, $E = 3n$, and $F = n + 1$, so $\chi = 1$ for every compatible n. Its sector action is dihedral, with group order $2n$. The image fixes $n = 8$, giving the accepted complete target

$$\left(16,\ \frac{a_1^8 + 5a_2^4 + 2a_4^2 + 4a_8 + 4a_1^2a_2^3}{16},\ 1\right).$$

The text does not fix eight rather than another allowed number of sectors. Thus a response can correctly derive $\chi = 1$ and a formula parameterized by n without determining the complete instance-specific target. Conversely, substituting eight without image evidence can produce the accepted tuple. The first Claude text-only response illustrates the latter event; its statement about the rendering is not justified by its input.

B.3 A LOOP CHANGES ONLY PART OF A MULTI-COMPONENT TARGET

GT-0382 asks for the first Betti number, the number of faces on the sphere, and the spanning-tree count of a connected seven-vertex planar multigraph. Its accepted tuple is $(10, 11, 620)$. A GPT first response omits a loop and uses 15 rather than 16 edges. The formulas $b_1 = E - V + 1$ and $F = E-V+2$ then produce nine and ten. Since a loop belongs to no spanning tree, the spanning-tree count remains 620. This response demonstrates why checking only one component would accept an incomplete solution.

B.4 PROOF OBLIGATIONS CAN DISTINGUISH IDENTICAL TERMINAL VALUES

GT-0019 asks for a hypermap genus and the minimal polynomial of a spectral radius, with a valid derivation and irreducibility argument. The accepted terminal pair is $(5, t^4 - 9t - 36)$. Claude's first original-image response incorrectly describes the strand permutation as the shift $x_k \mapsto x_{k+3}$ modulo 11. Although its terminal values agree, the wrong instance permutation does not prove the required statement about the depicted object. The complete-answer rubric therefore leaves it incorrect. For this task, a terminal-value-only comparison would not implement the question's stated requirements.

B.5 A COMPLETE FIRST-RESPONSE CENSUS WITHIN FOUR SELECTED CASES

To compare coarse checks with the specified targets, we read the complete first responses of all three models for GT-0411, GT-0382, and GT-0019 in O and GT-0514 in T. These four cases were selected after their behavior was known. Table 5 records the question-specific projection and complete-target outcomes. The comparison uses the final explicitly corrected answer, not an isolated matching number. In particular, Claude retracts its first GT-0411 factors and ends with $(8, 8, 8, 8, 8, 8)$; the stored parsed-answer field predates that correction. Both versions fail, so the official label is unaffected.

GT-0411's three models all give the correct lower group, but no complete upper group is correct. Gemini alone matches its order. GT-0382's GPT response gives $(9, 10, 620)$: the tree count matches while the other components fail. For GT-0019, GPT supplies both the correct pair and the required instance derivation and irreducibility proof. Claude's wrong permutation gives a matrix sharing the true characteristic polynomial, illustrating why that polynomial alone does not establish the depicted instance's derivation.

In GT-0514 T, GPT leaves n symbolic, Gemini asserts $n = 10$, and Claude guesses $n = 8$. All three derive $\chi = 1$, but only Claude matches the complete fixed target. Two responses report a concrete n, one correct and one wrong; GPT's unspecified parameter is recorded separately. Both Gemini and Claude attribute their chosen count to a rendering absent from their inputs. The companion records

Table 5: Post-hoc census of 12 first responses in four selected cases. Each row covers all three models. A projection keeps only the specified component or terminal values; the complete check follows the original question. These different projections do not define a common alternative benchmark metric. †: one of two concrete sector counts is correct; the third response leaves n unspecified. Complete hits use all three responses.

ID	Input	Projection checked	Projection hits	Complete hits
GT-0411	O	Upper group order, 82944	1/3	0/3
GT-0382	O	Tree count, 620	1/3	0/3
GT-0019	O	Terminal genus–polynomial pair	2/3	1/3
GT-0514	T	Sector count, $n = 8$	1/2†	1/3

the 12 source bindings, 33 located excerpts, extraction criteria, and exact arithmetic checks. These cases establish inspectable distinctions between target granularity and support; they do not estimate their dataset-wide frequency or calibrate the judge.

C Complete Per-Problem Results

Table 6 gives the four binary outcomes for each question, model, and evaluated input condition. The 42 Challenge-only questions have no T entries in the common analysis. The sum of the displayed positions is 984, including the one explicitly marked manual judgment. A zero denotes an incorrect label, not a missing response, except that the marked position has a user judgment without a complete local answer archive.

Table 6: All four-roll verdict vectors, in roll order (1=Correct, 0=Incorrect). O: original image; T: text only. T was evaluated only on Hard20. †: Claude GT-0560 T roll 2 uses a user-reported Incorrect adjudication without a complete archived response.

ID	Set	O: GPT	O: Gemini	O: Claude	T: GPT	T: Gemini	T: Claude
GT-0004	C	1111	1111	0001	—	—	—
GT-0008	C	1101	0000	1000	—	—	—
GT-0010	C	1111	0111	0000	—	—	—
GT-0012	C	0011	1101	0110	—	—	—
GT-0016	C	1001	1101	0000	—	—	—
GT-0018	C	1011	0000	0000	—	—	—
GT-0019	H	1001	0000	0000	0000	0000	0000
GT-0049	C	1111	0000	0000	—	—	—
GT-0050	H	0010	0000	0000	0000	0000	0000
GT-0061	H	0000	0000	0000	0000	0000	0000
GT-0070	C	1011	0010	0000	—	—	—
GT-0081	C	1010	0000	0000	—	—	—
GT-0084	C	0000	0000	0000	—	—	—
GT-0102	H	0000	0000	0000	0000	0000	0000
GT-0138	C	1100	0011	0000	—	—	—
GT-0163	C	0001	0011	0111	—	—	—
GT-0164	H	1000	0000	0000	0000	0000	0000
GT-0167	C	0010	0000	0000	—	—	—
GT-0168	C	1000	1111	1010	—	—	—
GT-0207	C	1111	1111	0000	—	—	—
GT-0208	C	0111	0000	0000	—	—	—
GT-0216	H	0000	0000	0000	0000	0000	0000
GT-0219	C	1111	0011	0000	—	—	—
GT-0230	C	0111	1111	1100	—	—	—
GT-0239	C	0100	0000	0000	—	—	—
GT-0244	H	0000	1100	0000	0000	0000	0000
GT-0248	H	0000	0000	0000	0000	0000	0000
GT-0255	C	1111	0100	0000	—	—	—
GT-0277	C	1110	1111	0101	—	—	—

ID	Set	O: GPT	O: Gemini	O: Claude	T: GPT	T: Gemini	T: Claude
GT-0282	C	0010	0000	0000	—	—	—
GT-0288	H	0000	0000	0000	0000	0000	0000
GT-0291	H	0000	0000	0000	0000	0000	0000
GT-0302	H	0000	0000	0000	0000	0000	0000
GT-0305	C	0000	1111	1111	—	—	—
GT-0317	C	1111	0101	0001	—	—	—
GT-0325	C	0010	0000	0000	—	—	—
GT-0361	H	0000	0000	0000	0000	0000	0000
GT-0372	C	0000	0001	0000	—	—	—
GT-0374	C	0000	0000	0000	—	—	—
GT-0382	H	0000	0000	0000	0000	0000	0000
GT-0385	C	0000	0000	0000	—	—	—
GT-0394	C	0001	1011	0000	—	—	—
GT-0395	H	0000	0000	0000	0000	0000	0000
GT-0402	C	0000	1111	0000	—	—	—
GT-0411	H	0000	0000	0000	0000	0000	0000
GT-0424	C	0000	1100	0000	—	—	—
GT-0426	C	0011	0000	0001	—	—	—
GT-0436	C	0110	1111	1100	—	—	—
GT-0437	C	0000	0000	0000	—	—	—
GT-0495	C	0000	1000	0110	—	—	—
GT-0503	C	0110	1001	0000	—	—	—
GT-0514	H	0000	0000	0000	0000	0000	1111
GT-0516	H	0000	0000	0000	0000	0000	0000
GT-0524	C	1000	0000	0000	—	—	—
GT-0526	C	1000	0010	0000	—	—	—
GT-0543	C	0000	0010	0000	—	—	—
GT-0547	C	1001	0000	0000	—	—	—
GT-0554	H	0000	0000	0000	0000	0000	0000
GT-0560	H	0000	0000	0100	0000	0000	0000†
GT-0572	C	0000	1011	0000	—	—	—
GT-0573	C	1111	1111	1110	—	—	—
GT-0585	H	0000	0000	0000	0000	0000	0000

D Content and Reference Evidence Index

Table 8 identifies each retained question and its requested target. All entries belong to Challenge62; H marks the nested Hard20 and C its complement. The output form describes the required answer rather than a complete taxonomy of the mathematics. "Group-valued" includes finitely generated abelian groups with a free summand, as in GT-0016 and GT-0018; it does not assert that every group is finite. The accompanying content cards bind these descriptions to the original question and image, the accepted target, and the corresponding verification record. Table 9 lists reference status and a summary of its verification basis; the editable evidence index additionally links the underlying records.

Table 7: Descriptive output forms across the nested evaluation sets. Counts concern questions, not responses.

Output form	Challenge	Hard
Integer tuple	19	6
Symbolic polynomial	14	6
Symbolic tuple	10	3
Integer	8	2
Group-valued	7	2
Structured list	4	1
Total	62	20

Table 8: Content index of all 62 current questions. H denotes Hard20; C denotes Challenge-only. Descriptions follow the saved content inventory.

ID	Set	Output	Requested target
GT-0004	C	Symbolic tuple	Compute alternating-group constituent dimension and character difference.
GT-0008	C	Group-valued	Determine nonunit Smith invariant factors of a reduced graph Laplacian.
GT-0010	C	Symbolic tuple	Compute ribbon-graph and checkerboard transition polynomials; embedding-sensitive graph invariants cross two leaves.
GT-0012	C	Integer tuple	Determine prime exponents of the spanning-tree count of an iterated line graph.
GT-0016	C	Group-valued	Determine free rank and torsion invariant factors of a graph Laplacian cokernel.
GT-0018	C	Group-valued	Determine the invariant-factor decomposition of an integral graph Laplacian cokernel.
GT-0019	H	Symbolic tuple	Compute hypermap genus and minimal polynomial of a block-transition spectral radius; coequal topology and spectral targets.
GT-0049	C	Symbolic polynomial	Determine the characteristic polynomial of a one-period exclusion transition matrix.
GT-0050	H	Integer	Count spanning trees of a line graph.
GT-0061	H	Integer tuple	Count maximal spanning forests, components and cycle-space dimension; forest enumeration is the main nontrivial target.
GT-0070	C	Integer tuple	Compute exponents of three graph critical groups.
GT-0081	C	Integer tuple	Compute graph cycle rank, bridge count and critical-group exponent; exponent is the main algebraic target.
GT-0084	C	Structured list	Count spanning trees of three threshold bipartite graphs.
GT-0102	H	Integer	Count spanning trees of the line graph of a reduced topological graph.
GT-0138	C	Integer tuple	Determine rotational symmetry order and reflection-axis count of a planar point set.
GT-0163	C	Symbolic polynomial	Compute reduced integral homology of a sphere quotient collapsing a cubical set.
GT-0164	H	Group-valued	Determine nonunit Smith factors of a spoke-rim line-graph reduced Laplacian.
GT-0167	C	Symbolic tuple	Compute endpoint-permutation cycle type, inversion number and order.
GT-0168	C	Integer	Determine the order of a diagram-derived graph critical group.
GT-0207	C	Integer tuple	Compute sum and product of spanning-tree counts of displayed tree line graphs.
GT-0208	C	Symbolic tuple	Compute cycle-index polynomials for a color-preserving cycle-automorphism group action.
GT-0216	H	Integer tuple	Count faces, internal edges, and side incidences of a planar cell decomposition.
GT-0219	C	Structured list	Recover ramification partitions of a genus one dessin.
GT-0230	C	Group-valued	Determine sandpile group invariant factors.
GT-0239	C	Integer	Count edge distinguished spanning trees of a weighted bundle cycle.
GT-0244	H	Symbolic tuple	Find the least selected word and sum its collection's inversion numbers.
GT-0248	H	Symbolic polynomial	Compute the chromatic polynomial of six disjoint graphs.
GT-0255	C	Symbolic polynomial	Enumerate spanning trees by internal activity.
GT-0277	C	Symbolic tuple	Compute cyclomatic number, directed simple cycle count, and adjacency characteristic polynomial.
GT-0282	C	Integer tuple	Count ribbon graph vertices, bands, and thickened surface boundary components.
GT-0288	H	Structured list	Count displayed closed curves, yellow triangular faces, and distinct fill hues by row.
GT-0291	H	Symbolic polynomial	Compute the adjacency characteristic polynomial of a disjoint clique union.
GT-0302	H	Integer tuple	Compute first Betti numbers of two finite stroke graphs.
GT-0305	C	Symbolic polynomial	Compute a recurrent label multigraph Tutte polynomial.
GT-0317	C	Integer tuple	Count spanning trees and identify the unique articulation vertex.

ID	Set	Output	Requested target
GT-0325	C	Integer	Count spanning trees of a color domain adjacency graph.
GT-0361	H	Integer tuple	Determine yellow induced subgraph components and vertex counts.
GT-0372	C	Integer tuple	Compute oriented chain boundary strata and coincident seam cancellation.
GT-0374	C	Group-valued	Determine a phase adjacency graph critical group.
GT-0382	H	Integer tuple	Compute spanning tree count alongside graph cycle rank and spherical face count.
GT-0385	C	Structured list	Compute a normalized rational nullspace basis of a weighted skew adjacency matrix.
GT-0394	C	Symbolic polynomial	Compute the weighted perfect matching generating polynomial.
GT-0395	H	Integer tuple	Compute graph cycle rank, complementary reduced component rank, and marked vertex count.
GT-0402	C	Symbolic tuple	Compute spanning tree count and the positive Laplacian spectrum.
GT-0411	H	Group-valued	Determine two critical groups in invariant factor form.
GT-0424	C	Symbolic polynomial	Compute the signed degree enumerator of the distinguished vertex across graph cells.
GT-0426	C	Symbolic tuple	Compute the independence polynomial of the interval inclusion tree and evaluate it at the pictured integer.
GT-0436	C	Symbolic polynomial	Sum matching generating polynomials over the ten displayed graphs.
GT-0437	C	Integer tuple	Count vertices, edges, and complementary faces of an embedded planar mesh.
GT-0495	C	Integer tuple	Compute augmented cycle ranks of three finite core graphs.
GT-0503	C	Symbolic polynomial	Compute the h-polynomial of a noncrossing-diagonal simplicial complex.
GT-0514	H	Symbolic tuple	Determine an automorphism-group order, sector-action cycle index, and cell-complex Euler characteristic.
GT-0516	H	Symbolic polynomial	Enumerate marked vertices by their visually assigned height strata.
GT-0524	C	Integer	Compute the Gram determinant of the chamber-set incidence matrix.
GT-0526	C	Integer	Count spanning trees of the central junction-and-trace graph.
GT-0543	C	Integer tuple	Compute rational incidence-matrix rank, with auxiliary blue-sector count.
GT-0547	C	Integer tuple	Count vertices, edges, and triangular simplices of the pictured graph's clique complex.
GT-0554	H	Symbolic polynomial	Compute the spanning-tree enumerator weighted by edges crossing a prescribed vertex partition.
GT-0560	H	Symbolic polynomial	Compute the expanded determinant polynomial of a signed monomial-weighted adjacency matrix.
GT-0572	C	Integer tuple	Determine periodic arrangement length and quotient-graph cycle rank and edge-orbit count.
GT-0573	C	Integer	Count spanning trees of the edge-labeled multigraph, distinguishing parallel tokens.
GT-0585	H	Symbolic polynomial	Construct the polynomial encoding every discrete output transition of a displayed decision-tree classifier.

Table 9: Reference status and verification basis for each current question. S: supported target without answer change; C: corrected target. Summaries describe the recorded checks; they do not imply independent human validation or automatic image verification.

ID	Status	Verification basis
GT-0004	S	Hook-length dimension agrees with a separate corner-removal recurrence.
GT-0008	S	Five boundary hubs give a wheel with 121 spanning trees; 251 minors give nonunit Smith factors (11,11).
GT-0010	S	Exact state-sum calculations support the target.
GT-0012	S	Source reconstruction and exact calculation support the retained target.
GT-0016	S	Contracting the marked endpoints gives a double-edge component, a 4-cycle, and a loop; the integer Smith form is (1,1,2,4,0,0,0).

ID	Status	Verification basis
GT-0018	S	The 57-vertex, 56-edge drawing reduces by 32 unimodular leaf operations to six 4-cycle blocks and a three-vertex path; the cokernel has free rank 7 and six factors of 4.
GT-0019	S	An eleven-arrow reconstruction, cycle and genus calculation, characteristic polynomial, and irreducibility check support the target.
GT-0049	S	Source reconstruction and exact calculation support the retained target.
GT-0050	C	Two image readings recover 32 edges; line-graph cofactors agree with rational elimination.
GT-0061	C	The omitted blue plaquette gives 26 plaquettes, eight components, three cycles, and 146 maximal spanning forests.
GT-0070	S	Source reconstruction and exact calculation support the retained target.
GT-0081	S	The trace has 12 core edges, 11 pendant tails, and 19 vertices. All cofactors give 165; the block critical groups give exponent 165.
GT-0084	S	Source reconstruction and exact calculation support the retained target.
GT-0102	C	Apparent crossings resolve into distinct junctions. The tree has 14 trivalent vertices and 16 leaves; line-graph cofactor and block product agree.
GT-0138	C	All 210 marker centers match the triangular-lattice annulus. Six rotations and six reflections preserve it; the outermost six-point layer excludes additional symmetries.
GT-0163	S	The red cubical set has two components, each with Euler characteristic 1 and no first homology, supporting the quotient polynomial.
GT-0164	S	Three qualifying spokes determine a complete graph on four vertices; 251 minors of its line-graph reduced Laplacian give factors (2,8,24).
GT-0167	S	Source reconstruction and exact calculation support the retained target.
GT-0168	S	An upper block with 30 trees joins a 4-cycle at an articulation; cofactor and enumeration of 495 candidate edge sets both give 120.
GT-0207	S	Source reconstruction and exact calculation support the retained target.
GT-0208	S	Source reconstruction and exact calculation support the retained target.
GT-0216	C	Two image computations find 102 blue sites, 102 bounded cells, and nine boundary segments per side; Euler and degree counts give (102,271,578).
GT-0219	S	Source reconstruction and exact calculation support the retained target.
GT-0230	S	Image reading and exact Smith normal form support the target.
GT-0239	C	Separate assistant source reconstructions and exact computations agree.
GT-0244	C	Eight blue dashed edges define the words; inversion enumeration preserves leading zeros and sums to 31.
GT-0248	S	The six diagrams are three 3-stars, two 4-cycles, and an 8-cycle. Edge-subset polynomial calculations agree with direct color enumeration.
GT-0255	S	Source reconstruction and exact calculation support the retained target.
GT-0277	S	Source reconstruction and exact calculation support the retained target.
GT-0282	S	The evidence records a source-specific mathematical or rule-based resolution.
GT-0288	C	Separate assistant source reconstructions and exact computations agree.
GT-0291	S	Pixel interpolation recovers all 13 bar heights; the disjoint-clique spectrum matches the complete target.
GT-0302	C	Visible-stroke and exterior-gap rules prohibit hidden bridges, giving zero pockets on the left and five on the right.
GT-0305	C	Exactly twice-occurring labels form a 6-cycle; labels occurring three times are excluded. Two image readings and all 64 edge-subset calculations agree.
GT-0317	C	The image agrees with the edge list; fourteen cofactors and tree enumeration correct the count to 3909.
GT-0325	S	The evidence records a source-specific mathematical or rule-based resolution.
GT-0361	C	The image has 64 yellow vertices. A 63-edge spanning-tree witness connects all four groups; full edge enumeration is not required.
GT-0372	C	The evidence records a source-specific mathematical or rule-based resolution.
GT-0374	S	A connected white corridor separates red and black. The region graph has two triangles sharing a vertex, with Smith factors (3,3).
GT-0382	C	Seven vertices and 16 edges, including the repeated arc and loop, give (10,11,620); all seven cofactors agree with tree enumeration.
GT-0385	C	Two image readings confirm the directed edge from 3 to 6. Exact rank is 2; the requested first two free-column basis vectors are well defined within the four-dimensional kernel.
GT-0394	S	Source reconstruction and exact calculation support the retained target.
GT-0395	C	Separate assistant image inspections and exact computation support the target.

ID	Status	Verification basis
GT-0402	S	The eleven-edge graph has six explicit orthogonal Laplacian eigenvectors with eigenvalues (0,3,3,5,5,6); all cofactors and enumeration of 462 candidate trees give 225.
GT-0411	C	Determinantal divisors give lower-prism factors (5,15), with 75 trees; the upper-graph factors are also supported.
GT-0424	C	The 21 signed cells and their degrees at vertex 6 sum to the corrected polynomial.
GT-0426	S	The evidence records a source-specific mathematical or rule-based resolution.
GT-0436	S	Source reconstruction and exact calculation support the retained target.
GT-0437	C	Five black tracks occur in each family; the prescribed count ignores colored curves.
GT-0495	S	Source reconstruction is documented in the reference evidence.
GT-0503	S	A separate source reconstruction and exact calculation support the target.
GT-0514	C	Separate assistant image inspections and exact computation support the target; the cell and symmetry formulas are detailed in Appendix B.2.
GT-0516	C	The five strata contain 1, 14, 24, 12, and 1 explicit node glyphs; multiple pixel thresholds and whole-glyph grouping agree.
GT-0524	S	Eighteen orange markers have matching colors. Row-set sizes (5,7,5,1) and their intersections give Gram determinant 93.
GT-0526	C	Twelve vertices and twenty edges give 24,000 spanning trees by all cofactors and edge-subset enumeration.
GT-0543	C	Seven clusters of eight glyphs give 56 vertices and 84 edges; the exact incidence rank is 55.
GT-0547	C	Separate assistant image inspections and exact computation support the target.
GT-0554	C	The corrected edge set removes two absent edges; exact weighted cofactors and full multiedge-tree enumeration agree.
GT-0560	C	Twelve arrows include the red arrow from 3 to 4 and no red loop at 3. Three determinant methods agree; see Appendix A.8 for the directional wording erratum.
GT-0572	S	A separate source reconstruction and exact calculation support the target.
GT-0573	S	A separate source reconstruction and exact calculation support the target.
GT-0585	C	All 48 paths identify no jump at 13 and 27; fifteen active boundary roots include four missing from the original reference.

E Reference-Correction Sensitivity

The fixed-response comparison in Table 10 measures the effect of reference correction and rejudgment on identical answer texts and fixed question memberships. The 53/136-question archived analysis sets, and the corresponding 52/134-question subsets without the earliest two corrected references, are separate from the quality-filtered Challenge62 and Hard20 membership reported in the main text. These sets include questions later excluded for unresolved quality or reference issues. The comparison therefore documents label sensitivity; it is not an additional current benchmark leaderboard.

F Additional Repeated-Success and Ablation Metrics

Figure 5 displays the distribution of correct responses per question; the complete metrics are given below. These reuse the same positions as the main results and add no scoring positions. Percentages are rounded only for presentation.

Table 10: Scores on fixed archived answer sets before and after reference correction and rejudgment. Rates are percentages and changes are percentage points. Answer texts and denominators remain unchanged within a row. Membership is held fixed for this sensitivity analysis.

Fixed membership	Model	Input	Slots	Saved	Corrected	Δ (pp)
Hard53	GPT	O	212	6.60	33.49	+26.89
Hard53	GPT	T	212	2.36	4.25	+1.89
Hard53	Gemini	O	212	2.36	34.91	+32.55
Hard53	Gemini	T	212	5.66	8.02	+2.36
Hard53	Claude	O	212	4.25	11.79	+7.55
Hard53	Claude	T	212	6.13	4.25	-1.89
Challenge136	Claude	O	544	11.40	16.54	+5.15
Hard52	GPT	O	208	5.29	32.69	+27.40
Hard52	GPT	T	208	0.48	2.40	+1.92
Hard52	Gemini	O	208	0.96	34.13	+33.17
Hard52	Gemini	T	208	3.85	6.25	+2.40
Hard52	Claude	O	208	2.40	10.10	+7.69
Hard52	Claude	T	208	4.33	2.40	-1.92
Challenge134	Claude	O	536	10.26	15.49	+5.22

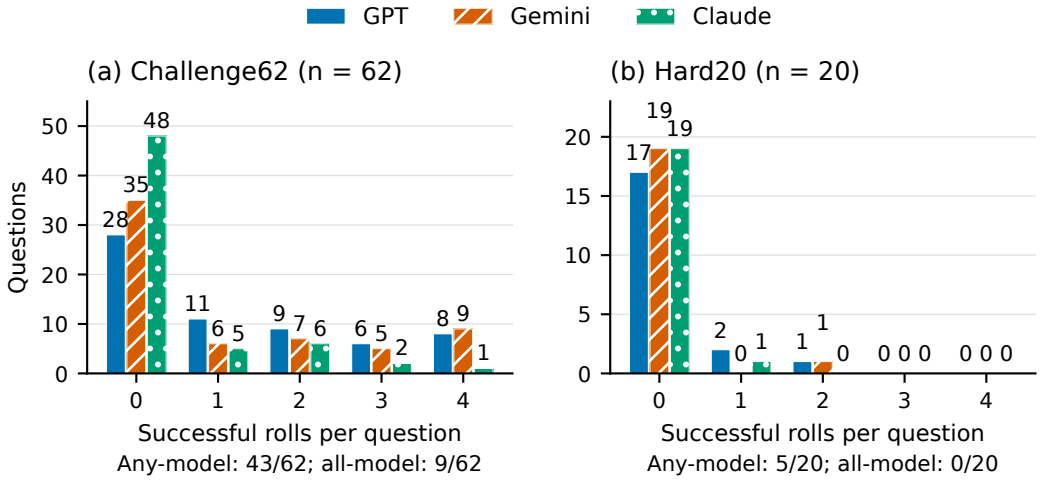

Figure 5: Distribution of the number of correct responses per problem. The same four responses underlie all accuracy and repeated-success metrics. Challenge62 contains both recoverable and persistent failures; the distribution on Hard20 is concentrated at zero successes.

Table 11: Hard20 scores under O and T. Rates and intervals are percentages. Correct counts refer to 80 response positions per model and condition; Acc@1, Pass@4, and Stable@4 use 20 questions.

Model	Input	Correct	Acc@1	MeanAcc	Pass@4	Stable@4
GPT-5.6 Luna	O	4/80	10.00 [2.8, 30.1]	5.00 [0.0, 11.2]	15.00 [5.2, 36.0]	0.00 [0.0, 16.1]
GPT-5.6 Luna	T	0/80	0.00 [0.0, 16.1]	0.00 [0.0, 0.0]	0.00 [0.0, 16.1]	0.00 [0.0, 16.1]
Gemini 3.6 Flash	O	2/80	5.00 [0.9, 23.6]	2.50 [0.0, 7.5]	5.00 [0.9, 23.6]	0.00 [0.0, 16.1]
Gemini 3.6 Flash	T	0/80	0.00 [0.0, 16.1]	0.00 [0.0, 0.0]	0.00 [0.0, 16.1]	0.00 [0.0, 16.1]
Claude Opus 4.8	O	1/80	0.00 [0.0, 16.1]	1.25 [0.0, 3.8]	5.00 [0.9, 23.6]	0.00 [0.0, 16.1]
Claude Opus 4.8	T	4/80	5.00 [0.9, 23.6]	5.00 [0.0, 15.0]	5.00 [0.9, 23.6]	5.00 [0.9, 23.6]

Table 12: Question counts by number of successful responses among four attempts. Challenge62 includes Hard20; the two rows for a model must not be added.

Subset	Model	0	1	2	3	4	Recovered after roll 1
Challenge62	GPT	28	11	9	6	8	13
	Gemini	35	6	7	5	9	10
	Claude	48	5	6	2	1	8
Hard20	GPT	17	2	1	0	0	1
	Gemini	19	0	1	0	0	0
	Claude	19	1	0	0	0	1
Challenge-only42	GPT	11	9	8	6	8	12
	Gemini	16	6	6	5	9	10
	Claude	29	4	6	2	1	7

G Uniform Exclusion of the Conflicting Item

This sensitivity removes GT-0560 for every model and condition, addressing its contradictory directional wording and the missing complete Claude T response on a common population. "Challenge61" and "Hard19" below denote this exclusion analysis only; the declared benchmark memberships remain Challenge62 and Hard20. The retained positions are all complete: $61 \times 3 \times 4 + 19 \times 3 \times 4 = 960$, comprising 180 correct and 780 incorrect labels. The reduction of 24 positions removes one correct response and 23 incorrect positions, including the manual decision. No remaining response, reference, or label is changed.

Table 13: Uniform GT-0560 exclusion. All rates and intervals are percentages. MeanAcc intervals resample questions; binary metric intervals are Wilson intervals. C61 and H19 denote Challenge61 and Hard19. Each row retains four responses per question.

Set	Model	Input	Correct	Acc@1	MeanAcc	Pass@4	Stable@4
C61	GPT	O	79/244	34.43 [23.7, 47.0]	32.38 [23.8, 41.4]	55.74 [43.3, 67.5]	13.11 [6.8, 23.8]
C61	Gemini	O	71/244	27.87 [18.2, 40.2]	29.10 [19.7, 38.9]	44.26 [32.5, 56.7]	14.75 [8.0, 25.7]
C61	Claude	O	26/244	9.84 [4.6, 19.8]	10.66 [5.3, 16.8]	21.31 [12.9, 33.1]	1.64 [0.3, 8.7]
H19	GPT	O	4/76	10.53 [2.9, 31.4]	5.26 [0.0, 11.8]	15.79 [5.5, 37.6]	0.00 [0.0, 16.8]
H19	Gemini	O	2/76	5.26 [0.9, 24.6]	2.63 [0.0, 7.9]	5.26 [0.9, 24.6]	0.00 [0.0, 16.8]
H19	Claude	O	0/76	0.00 [0.0, 16.8]	0.00 [0.0, 0.0]	0.00 [0.0, 16.8]	0.00 [0.0, 16.8]
H19	GPT	T	0/76	0.00 [0.0, 16.8]	0.00 [0.0, 0.0]	0.00 [0.0, 16.8]	0.00 [0.0, 16.8]
H19	Gemini	T	0/76	0.00 [0.0, 16.8]	0.00 [0.0, 0.0]	0.00 [0.0, 16.8]	0.00 [0.0, 16.8]
H19	Claude	T	4/76	5.26 [0.9, 24.6]	5.26 [0.0, 15.8]	5.26 [0.9, 24.6]	5.26 [0.9, 24.6]

Table 14: Paired MeanAcc differences after uniform exclusion, in percentage points. All pairs use the same retained questions on both sides, 100,000 bootstrap replicates, and seed 20260910.

Population	Comparison	Difference	95% CI
Challenge61	GPT O minus Gemini O	+3.28	[-6.97, 13.52]
Challenge61	GPT O minus Claude O	+21.72	[12.30, 31.56]
Challenge61	Gemini O minus Claude O	+18.44	[11.07, 26.23]
Hard19	GPT O minus Gemini O	+2.63	[-5.26, 10.53]
Hard19	GPT O minus Claude O	+5.26	[0.00, 11.84]
Hard19	Gemini O minus Claude O	+2.63	[0.00, 7.89]
Hard19	GPT T minus GPT O	-5.26	[-11.84, 0.00]
Hard19	Gemini T minus Gemini O	-2.63	[-7.89, 0.00]
Hard19	Claude T minus Claude O	+5.26	[0.00, 15.79]

The O success union changes from 43/62 to 42/61 and from 5/20 to 4/19. Claude's only Hard O success is removed; its four T successes on GT-0514 remain. All three Hard19 paired T-minus-O intervals still include zero, and the all-zero O or T means have degenerate empirical bootstrap intervals. This analysis shows the dependence of individual scores on the item while preserving the broader pattern of low Hard success. It does not determine how any model would respond to a corrected prompt.

Behavior sensitivity removes all six GT-0560 first responses without recoding the other annotations. Among 57 O responses, 49 have an instance-statement conflict, two a modeling conflict, one a condition conflict, one a computational conflict, one an unresolved error, and three consistent target and checked reasoning. Among 57 T responses, 22 are justified refusals, four general methods, nine final guesses, two empty-object interpretations, and 20 unsupported instance answers. Explicit

absent-image attribution occurs in eight, with six ambiguous and 43 negative. These are the original model-assisted categories on a reduced population, not a new independent annotation study.

H PER-QUESTION TEXT-SUFFICIENCY ARGUMENTS

Table 15 records the omitted facts, compatible alternatives or local changes, and affected target components for Hard20. A change to one required component leaves the complete target undetermined, even when other components follow from text. These model-assisted arguments range from explicit small alternatives to changes that hold other structure fixed; they are not 20 complete replacement images or independently certified structures.

Three qualifications remain. GT-0216 lacks explicit alternative cell complexes satisfying every degree condition. GT-0554 does not specify the added edge's endpoints or expanded polynomial. GT-0585 changes the active boundary, but two new factors would require an interior singleton. The all-20 audit finding retains these evidential limits. Original annotations, source lines, and scope notes accompany the table in the companion.

Table 15: Existing Hard20 text-sufficiency arguments. Alternatives describe compatible changes to facts omitted from the text; they are not a collection of complete replacement images. The qualifications below and in the text delimit the evidence.

ID and missing fact	Existing alternative or local change	Changed target and qualification
GT-0019 Endpoint permutation.	On eleven darts, $\sigma = r$ versus $\sigma = r^{-1}$.	Genus: 5 versus 0.
GT-0050 Marker count and adjacency.	Use G=K_3 versus G=K_2.	Line-graph tree count: 3 versus 1.
GT-0061 Filled torus plaquettes.	On $\mathbb{Z}_{10} \times \mathbb{Z}_{10}$, use {(0,0)} versus {(0,0),(2,2)}.	Tuple: (1,0,1) versus (2,0,1).
GT-0102 Stroke branches and endpoints.	After the prescribed deletion/suppression, retain a path versus a tree with one degree-three branch.	Line-graph tree count: 1 versus 3.
GT-0164 Spoke count N.	Use the specified construction with N=3 versus N=4.	Nonunit Smith factors: (2,8,24) versus (33,264).
GT-0216 Cell count N; boundary count b.	A generic cell complex versus a local generic refinement increasing N.	N changes; $E = 3N + 1$ and $E_{\mathrm{int}} = 3N + 1 - b$. Complete compatible cell pair not supplied.
GT-0244 Dashed edges; fills and labels.	Relabel the unfilled endpoint of one dashed edge from 01325 to 13025.	Tuple: (1,01325,1) versus (1,13025,3).
GT-0248 Six source adjacencies.	Replace one selected P_4 by C_4; hold the other five graphs fixed.	Chromatic factor: $q(q-1)^3$ versus $(q-1)^4 + (q-1)$.
GT-0288 Panel curves, faces and hues.	Change the upper-right face colors from yellow/green to yellow/yellow.	Distinct-hue component p1: 2 versus 1.
GT-0291 Thirteen bar heights.	Change one bar from height 6 to 8, away from half-integers; hold other bars fixed.	One characteristic-polynomial factor changes from K_6 to K_8.
GT-0302 Open versus closed returns.	Close a genuine gap in one return; hold the rest fixed.	Local cycle contribution: 0 versus 1; a Betti component changes.
GT-0361 Yellow nodes and adjacency.	Two yellow nodes with versus without an edge.	Tuple: (1,2,2) versus (2,1,2).
GT-0382 Arc endpoints, loops and multiplicities.	Add a loop at f to a compatible seven-vertex plane multigraph, preserving the other arcs.	Cycle rank and face count each increase by 1; tree count is unchanged.
GT-0395 Incidences and marked vertices.	Add one degree-two arrowhead mark on an existing arc.	Arrowhead count a increases by 1; r and s are unchanged.
GT-0411 Two graph adjacencies.	Use C_4 versus K_4 for the upper graph; hold the lower graph fixed.	Upper critical group: $\mathbb{Z}/4\mathbb{Z}$ versus $(\mathbb{Z}/4\mathbb{Z})^2$.

ID and missing fact	Existing alternative or local change	Changed target and qualification
GT-0514 Sector count n.	Use the central polygon and triangular fan with n=6 versus n=8.	Group order: 12 versus 16; Euler characteristic remains 1.
GT-0516 Intermediate-stratum node counts.	Add one visible node mark at height 1, preserving the five strata.	Coefficient of t increases by 1.
GT-0554 Source edge lists.	Take A=B=P_6; add a cross-I/J edge in one source, preserving the stated matching.	Weighted tree polynomial changes. Added edge and expansion unspecified.
GT-0560 Arrow endpoints and labels.	Change the exponent pair (r,s) on one loop; hold other arrow data fixed.	Trace of M changes, hence the coefficient of z in $\det(I - zM)$.
GT-0585 Thresholds and terminal classes.	From all leaves labeled no jump, relabel one reachable singleton as jump.	Active boundary and transition polynomial change. Two factors require an interior singleton.