TopoEval: A Comprehensive Benchmark for Topological Reasoning in Foundation Models

Yinzhu Cheng[1,2,†] **Yihan Meng**[1,2,†] **Haihua Xie**[2,*] **Mingming Sun**[2,*]
[1]Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China
[2]Beijing Institute of Mathematical Sciences and Applications, Beijing 101408, China
[†]Equal contribution. [*]Corresponding authors.
{chengyinzhu,mengyihan,haihua.xie,sunmingming}@bimsa.cn

Abstract

Topological reasoning—the ability to identify structural invariants under continuous deformations—is a cornerstone of human visual cognition, yet it remains under-explored in modern AI systems. Existing benchmarks provide limited coverage of mathematical topology as a discipline and rarely probe the depth of topological reasoning. To bridge this gap, we introduce TopoEval, an 800-item benchmark suite for topological reasoning. TopoEval centers on a 400-item answer-based Core track and includes a 400-item Proof track for reference-grounded informal proof assessment. The suite spans four major topology branches, twelve fine-grained subfields, and three difficulty levels. On Core, we investigate model scaling, reasoning depth, prompt strategies, image dependency, and academic tiers. The strongest Core configuration reaches 73.0% overall and 54.1% on all 85 K12 items. For context, topology-trained graduate-student solvers reach 96.3% under the audited protocol in Appendix B.2. On Proof, pass-plus-borderline rates are low across the selected conditions, indicating difficulty in producing reference-aligned natural-language topology proofs. TopoEval profiles these capabilities across academic level, image dependency, and proof quality.

1 Introduction

Topology, a fundamental branch of modern mathematics, studies properties of **geometric objects that remain invariant under continuous deformations** (Munkres, 2000), such as stretching and bending without tearing or gluing. For instance, although a coffee mug and a donut differ substantially in geometric appearance, they are topologically equivalent because both contain exactly one "hole" (Figure 2) (Lee, 2012). Reasoning about structural invariants is a hallmark of human visual cognition. Cognitive studies indicate that this form of intuition, including the ability to identify connectivity and holes, emerges as early as infancy (Chien et al., 2012). As foundation models become central components of modern AI systems, a fundamental question arises: **Do current state-of-the-art foundation models exhibit topological understanding comparable to that of humans?**

Recent progress in foundation models has led to the emergence of increasingly capable systems, including open-weight Qwen3 (Qwen Team, 2025) and LLaMA-4 (Meta AI, 2025), as well as proprietary GPT-5.2 (OpenAI, 2025b) and Gemini-3 (Hassabis et al., 2025). These modern models combine advanced reasoning mechanisms with multimodal perception. Enhanced reasoning capabilities, including explicit chain-of-thought reasoning, self-consistency, and search-based inference strategies (e.g., Monte Carlo Tree Search), have led to notable improvements in complex logical and multi-step reasoning (Sui et al., 2025; Xie et al., 2024). Advances in large-scale cross-modal alignment, including vision–language pretraining and instruction tuning over multimodal data, have substantially strengthened visual understanding (Yin et al., 2024; Liu et al., 2023). As a result, **visual reasoning has emerged as a critical testbed** for evaluating how effectively abstract reasoning and perceptual grounding are integrated within a single model (Lu et al., 2024). Many topology problems require this kind of visual-structural reasoning.

General-purpose datasets such as MathVista (Lu et al., 2024), MMMU (Yue et al., 2024), and MathVision (Wang et al., 2024) provide broad multimodal coverage, while TopoBench (Maniparambil et al., 2026) tests six grid-puzzle families. TopoEval contributes a domain-deep evaluation of mathematical topology spanning visual intuition, academic concepts, and professional proofs (Appendix A.2).

To address this gap, we introduce **TopoEval**, **an 800-item benchmark suite specifically designed to evaluate topological reasoning in foundation models**. Figure 1 illustrates representative

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 31391–31421
October 24-29, 2026 ©2026 Association for Computational Linguistics

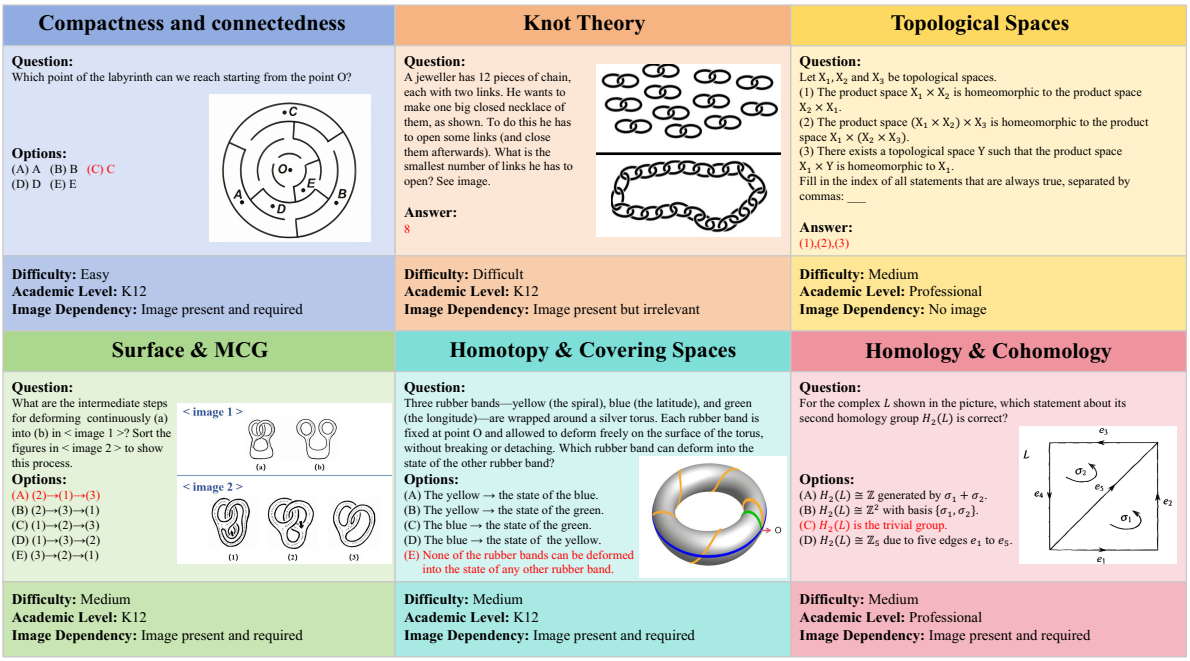


Figure 1: **Example problem cards from TopoEval.** Correct answers are highlighted in red. Displayed examples emphasize K12-level problems for accessibility; Section 3.1 reports the full-suite distribution.



Figure 2: **Topological Equivalence**: Both a coffee cup and a donut have one hole and can be mutually transformed via continuous deformation without tearing or gluing. They are regarded as topologically equivalent.

items across subfields, academic levels, and image-dependency categories. **TopoEval-Core** contains 400 multiple-choice or fill-in-the-blank items and reports final-answer accuracy. **TopoEval-Proof** contains 400 professional proof tasks, with candidate generation followed by reference-grounded auditing that reports pass, borderline, and fail rates. Across the suite, items cover point-set, geometric, algebraic, and differential topology and twelve fine-grained subfields. Core spans K12 to professional training and includes image-dependency annotations for visual-grounding analysis.

Using TopoEval-Core, we evaluate open-weight and proprietary models across model scale, reasoning effort, and prompting strategies. The strongest configuration reaches 73.0% overall and 54.1% on all 85 K12 items. The human reference reaches 96.3% under the audited protocol in Appendix B.2. This comparison measures model headroom on the K12 material. K12 items emphasize knots, holes, connectivity, and continuous deformation, and model performance is also markedly weaker when solving requires visual evidence. Together, these results locate a persistent difficulty in intuitive and perceptual topological reasoning alongside the challenge of professional knowledge.

Our main contributions are threefold:

- **Domain-Deep Benchmark Design:** We organize 800 items across four branches, twelve subfields, and K12-to-professional levels, with Core answer evaluation and reference-grounded Proof assessment.

- **Systematic Evaluation:** We compare representative foundation models across scale, reasoning effort, prompting, image dependency, and academic tier, and report persistent gaps on K12 and image-required items.

- **Analysis and Insights:** Fine-grained results and case studies identify recurring errors in visual structure tracking, invariant selection, and proof closure.

2 Related Work

2.1 From Training-Time Scaling to Inference-Time Reasoning

Long-standing improvements in foundation models have largely followed *classic scaling laws*: by jointly increasing model parameters, training data, and compute, models achieve systematic performance gains across a wide range of tasks (Kaplan et al., 2020; Hoffmann et al., 2022). At larger scales, the cost of high-quality training data rises and marginal returns diminish (Villalobos et al., 2022; Chen et al., 2025), creating practical bottlenecks for training-time scaling.

Inference-time scaling allocates more computation during inference (Snell et al., 2024). Reasoning-augmented frontier systems, including xAI's Grok 4.1 and Anthropic's Claude Opus 4.5, exemplify this direction (xAI, 2025b; Anthropic, 2025a). Their inference capabilities are commonly supported by supervised fine-tuning and reinforcement-learning-based post-training, which elicit multi-step reasoning behavior and search over reasoning paths (Ouyang et al., 2022; Bai et al., 2022).

Additional inference compute has practical limits. Model thought traces can contain redundant reasoning, increasing computational cost and sometimes reducing accuracy on intuition-heavy problems (Snell et al., 2024; Sui et al., 2025).

Many problems in TopoEval benefit from geometric intuition for efficient reasoning-path search, making the benchmark well suited for testing whether foundation models gain structural understanding beyond increases in reasoning budget.

2.2 Visual Grounding and Hallucination in Multimodal Reasoning

Multimodal large language models (MLLMs) typically learn shared representations of vision and language through joint pre-training on large-scale image–text pairs (Radford et al., 2021; Jia et al., 2021). In this paradigm, models often employ frozen visual encoders aligned with large language models via cross-modal projection layers or cross-attention mechanisms (Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2023). Within this framework, visual grounding and hallucination have emerged as central challenges (Peng et al., 2023; Chen et al., 2023; Liu et al., 2024). Prior studies indicate that hallucination is particularly pronounced when models fail to exploit visual evidence and fall back on language priors acquired during pre-training (Li et al., 2023b; Liu et al., 2024).

Topological reasoning amplifies these visual-grounding challenges (Munkres, 2000). Problems involving mazes, knots, or continuous deformations (Adams, 2004) center on irregular curves and complex connectivity. Small local visual discrepancies can change entanglement or connectivity and lead to different topological conclusions.

TopoEval-Core measures this demand through image-dependency annotations that distinguish text-only items, items with non-essential images, and items whose solution requires visual evidence. Of the 400 Core items, 83 are text-only, 57 contain a non-essential image, and 260 require visual evidence. Image-required items comprise 87.06% of the K12 split and 59.05% of the professional split. This dimension supports separate analysis of textual reasoning and visual grounding.

Table 1: **Dataset statistics of the TopoEval suite.** TopoEval-Core and TopoEval-Proof are separate tracks and are reported with different metrics.

TopoEval-Core: answer-based track	
Items	400
Format	MCQ 324; FIB 76
Level	K12 85; Professional 315
Image	No image 83; present-only 57; required 260
Metric	Answer accuracy
TopoEval-Proof: proof-auditing track	
Items	400
Format	Informal proof audit
Level	Professional 400
Image	Text-only 400
Metric	Reference-grounded proof audit

3 Benchmark Suite

3.1 TopoEval Overview

TopoEval is a curated 800-item benchmark suite for evaluating topological reasoning in foundation models. It contains TopoEval-Core, a 400-item answer-based track adapted from textbooks, lecture notes, and competition sources, together with TopoEval-Proof, a separate 400-item proof-auditing track built from professional-level topology statements. Table 1 summarizes the two-track suite statistics.

For TopoEval-Core, we categorize image dependency into three levels: no image, image present but not required for solving the problem, and image

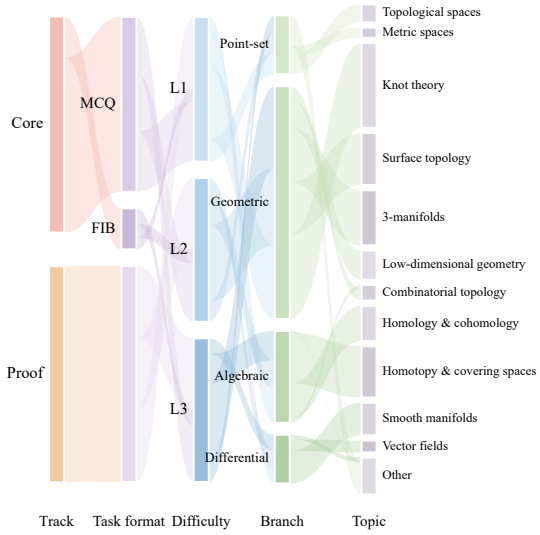


Figure 3: **Suite-level metadata flow.** Sankey-style view of the 800-item suite. Columns show track, task format, difficulty, branch, and topic. Proof difficulty uses the empirical tertiles described in the text. Topic labels are defined in Table 7.

present and required for reaching the correct solution. This annotation supports the visual-grounding analysis in Section 4.

Each problem is categorized into a hierarchical taxonomy. At the top level, TopoEval spans four major branches—*Point-set*, *Algebraic*, *Geometric*, and *Differential*. Figure 3 visualizes the suite-level metadata flow across track, task format, difficulty, branch, and topic. For Proof, L1–L3 are empirical tertiles of mean rubric score across the ten conditions in Table 12. The rightmost topic labels correspond to the twelve fine-grained topological subfields defined in Table 7.

3.2 Data Collection

We collaborated with researchers specializing in topology to identify reliable sources for TopoEval, including 14 textbooks/monographs, two sets of course lecture notes, and two competition or assessment systems. Concretely, TopoEval draws problems from standard topology textbooks and monographs (e.g., *Topology* (Munkres, 2000) and *Elements of Algebraic Topology* (Munkres, 1984)), course lecture notes (e.g., *Lecture Notes on Point-Set Topology* (Xiong, 2011) and *Lecture Notes on Basic Topology* (You, 1997)), as well as competition-style assessments that emphasize intuitive and visual understanding (e.g., *Math Kangaroo* (Math Kangaroo Association, 2025) and *Caribou Contests* (Caribou Mathematics Contest, 2025)).

During data collection, we first apply GPT-5.1 (OpenAI, 2025a) for OCR-based text extraction from scanned pages. We then use foundation models for automated consistency checks, filtering out items with malformed notation or internal contradictions. Finally, trained human reviewers remove problems with insufficient topological content or unclear visual evidence. Suitable Core items are adapted into evaluable multiple-choice or fill-in formats following the curation protocol in Sections 3.3–3.4, while Proof items retain natural-language proof obligations under the protocol summarized in Appendix D.

3.3 Data Adaptation and Curation

Many topology problems are naturally stated as proof tasks. TopoEval retains this form in the Proof track and uses answer-based formats in Core for reproducible scoring. Interactive theorem provers such as Lean and Isabelle provide machine-checkable verification for formalized statements (de Moura et al., 2015; Nipkow et al., 2002); TopoEval-Proof evaluates informal, visual, and advanced arguments through reference-grounded auditing (Avigad, 2018; Azerbayev et al., 2023; Poiroux et al., 2025). The two tracks capture final-answer behavior and reference-aligned natural-language proof quality, respectively.

For TopoEval-Core, we systematically adapt suitable problems into multiple-choice or fill-in-the-blank formats to ensure reproducible answer accuracy. In objectively scored visual-math settings, prior benchmarks commonly use structured answer extraction and normalization to support automatic evaluation (Lu et al., 2024; Wang et al., 2024). Reformulation also reduces exact textual overlap with source problems and solution corpora (Kapoor and Narayanan, 2023; Ramos et al., 2025). Core measures final-answer behavior through a controlled scoring interface. TopoEval-Proof supplies proof-level evaluation through rubric judgments against private reference-only fields. The candidate-generation prompt is given in Appendix E.2, and the paper-facing audit rubric is summarized in Appendix D.

During the adaptation stage, we use an LLM-assisted annotation and adaptation system. The system produces structured annotations covering candidate answers, fine-grained topological subfields, image dependency, academic level, and difficulty estimates. All outputs are subsequently reviewed by topology-trained annotators under the protocol

below.

3.4 Construction and Quality Control

Structured records. The annotation environment converts textbook pages, lecture notes, competition problems, images, and adapted variants into a common record. Each item stores the question, task type, options when applicable, answer, subfield, source, image path, image-dependency label, academic level, and difficulty. Core answer prompts, evaluation inputs, spreadsheets, and image bundles are generated from these records, keeping the item text, answer, metadata, and visual asset aligned across the pipeline. Proof records use the same taxonomy fields together with candidate-facing proof statements and task instructions.

Annotator protocol and adjudication. Annotators received shared instructions covering the record schema, the Core–Proof distinction, admissible transformations into multiple-choice or fill-in-the-blank formats, answer verification, image-dependency labeling, source traceability, and preservation of mathematical substance. Adaptations use the smallest rewrite needed for reproducible evaluation. Items with uncertain answers, diagram dependency, or subfield assignment are flagged during review.

Each retained item was checked by reviewers with topology training. Their review covered mathematical correctness, answer validity, task admissibility, subfield and image-dependency labels, and Proof release fields. Disagreements on correctness or metadata were discussed in a review meeting, with a third topology-trained reviewer invited to adjudicate. Ambiguous items were revised, deferred for further checking, or excluded. Annotation and review were carried out by project members and collaborators as part of the research effort.

Role-specific assistance. The environment assigns narrow roles to LLM assistance. OCR extracts text from screenshots and scans; adaptation converts partial drafts into the target record schema; review checks clarity, structural compliance, answer consistency, and MathJax compatibility; LaTeX repair uses rendering feedback to correct notation; and translation preserves mathematical notation across English and Chinese. Generation, review, translation, and repair outputs remain separate in the audit trail, so annotators can inspect the source of each proposed change.

Within this workflow, generated records are parsed against the expected schema. Missing or malformed fields trigger a format-repair request using the previous response and required contract. The reviewer returns a structured pass/fail judgment with issue lists and revision guidance. Failed reviews enter a capped sequence of revision rounds and then proceed to human inspection. Mathematical notation is rendered with MathJax, and parser errors are sent to the LaTeX repair role together with the original snippet. These checks cover question text, options, and answer fields.

Image and export integrity. For image-bearing records, the interface supports image preview, renaming, path normalization, and joint export of the spreadsheet and image bundle. This preserves the association between a question and its figure during evaluation and retains image dependency as an analysis field. Figure 4 summarizes the construction path from source intake to export.

Human control and provenance. Across these stages, annotators can edit records and inspect model configurations, intermediate responses, review feedback, and render reports. Automatic checks provide structured inputs to human review. Final answers, ambiguous cases, metadata assignments, and exported records receive manual approval. Source and version metadata remain attached to the retained items, supporting later audits and versioned evaluation runs.

4 Experiments

This section evaluates representative foundation models on TopoEval-Core. All accuracy values and model rankings refer to Core; single-run and voting conditions are labeled where reported. We analyze overall and subfield accuracy, parameter and inference-time scaling, image dependency, academic tier, Chain-of-Thought prompting, and multi-sample aggregation (Wei et al., 2022; Snell et al., 2024). We also report selected TopoEval-Proof conditions on their pass, borderline, and fail scale.

4.1 Experimental Setup

Following prior work on large-scale foundation model evaluation, we select representative models using established model comparison benchmarks and public leaderboards (Lu et al., 2024; OpenRouter, 2024). Our open-weight set covers several major foundation model families, including the

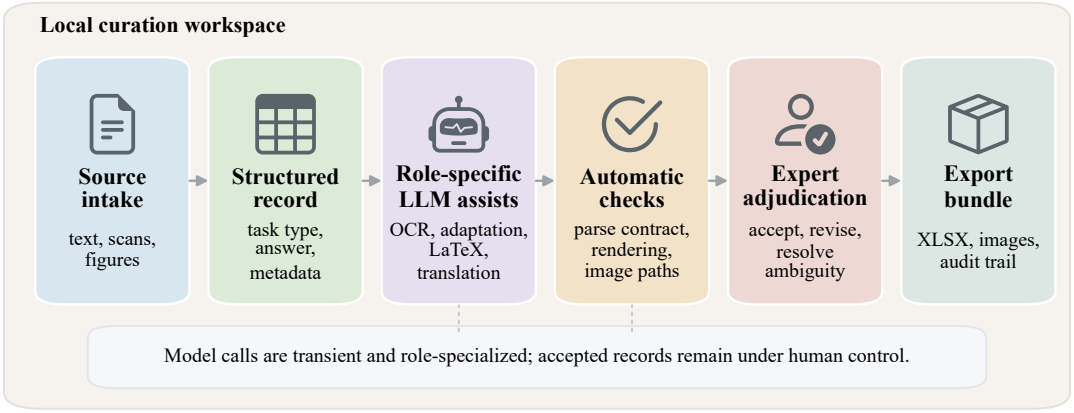
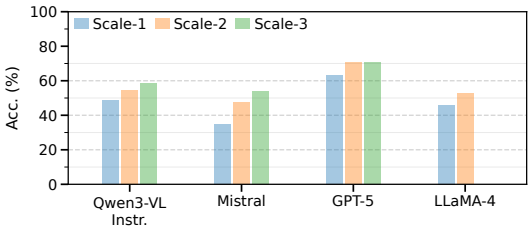
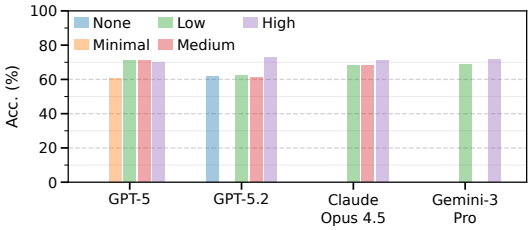


Figure 4: **Overview of the LLM-assisted annotation and adaptation system.** Heterogeneous source material is converted into structured records through role-specific LLM assistance, automatic parse/render checks, and explicit expert adjudication before export. Final benchmark records remain under human control.



(a) Accuracy under different parameter scales.



(b) Accuracy under different inference-time reasoning effort levels.

Figure 5: **Scaling trends from training-time capacity to inference-time computation.** Table 4 defines the model scale mappings.

Qwen3-VL family (Qwen Team, 2025), LLaMA-4 variants (Meta AI, 2025), GLM-4.6V (Z.ai, 2025), ERNIE-4.5-VL (Baidu ERNIE Team, 2025), and the Mistral/Ministral series (Mistral AI, 2025). Our proprietary set includes frontier systems from major providers, including GPT-5 and GPT-5.2 (OpenAI, 2025b), the Gemini-3 series (Hassabis et al., 2025), the Claude-4.5 series (Anthropic, 2025b), and Grok-4 and Grok-4.1 (xAI, 2025a).

Unless otherwise specified, all Core models are evaluated zero-shot with temperature 0 and Top-p 0.75. GPT-5-nano (OpenAI, 2025b) provides unified answer extraction, normalization, and correctness verification based on the reported final answer. The answering and judge interfaces use strict JSON formats; in 100 inspected Core judgments, we found one extraction-related error. Appendix B reports protocol clarifications, and Appendix B.3 gives Wilson intervals for key rates. Core experiments are conducted through the OpenRouter API (OpenRouter, 2024), which provides a unified interface for accessing both open-weight and proprietary models. Candidate-facing prompts and the paper-facing scoring schemas are reported in Appendix E.2.

4.2 Experimental Results

Parameter-side vs. inference-time scaling. Figures 5a and 5b compare parameter-side scaling and inference-time scaling. Across the plotted model families, both increasing model size and allocating more computation at inference time can improve performance. The gains plateau for some families, including GPT-5, and higher reasoning effort produces uneven changes. Three-sample voting is similarly variable, ranging from a 1.5-point gain to a 1.5-point decline in representative runs (Table 8). These patterns are consistent with additional test-time computation being most useful once a model has formed an adequate visual and structural representation.

Overall results. Figure 6 summarizes performance across fine-grained topology subfields, and Table 2 reports detailed accuracy results across all 12 subfields, whose sample sizes range from 12

Table 2: **Performance comparison across 12 fine-grained topology subfields.** We report accuracy (%) for representative models and selected scale anchors. Red indicates the best performance, and Blue indicates the second best. Parenthetical qualifiers indicate inference-time reasoning effort or voting where shown. See Table 7 for the definitions of the 12 fine-grained subfields. Results for the remaining models are reported in Table 11.

Model	Knot	Surf	3M	Hom	Hty	LDG	VFld	TopSp	SmMfd	Comb	Metric	Other
Proprietary Models												
GPT-5 (minimal)	51.5	63.2	58.8	56.2	85.7	72.0	57.1	60.0	68.4	60.0	91.7	50.7
GPT-5 (high)	54.5	73.7	55.9	90.6	71.4	84.0	66.7	80.0	89.5	80.0	100.0	59.2
GPT-5 Mini	57.6	68.4	61.8	84.4	71.4	84.0	57.1	95.0	94.7	73.3	100.0	63.4
GPT-5.2 (none)	52.5	64.9	52.9	75.0	67.9	76.0	57.1	75.0	73.7	63.3	97.1	56.3
GPT-5.2 (high)	57.6	73.7	70.6	93.8	75.0	88.0	66.7	95.0	89.5	83.3	100.0	69.0
Gemini-3 Flash	53.5	75.4	64.7	93.8	75.0	92.0	71.4	95.0	89.5	80.0	100.0	66.2
Gemini-3 Pro (high)	54.5	75.4	64.7	93.8	78.6	92.0	71.4	95.0	94.7	80.0	100.0	66.2
Claude Opus 4.5 (low)	48.5	64.9	58.8	93.8	67.9	80.0	66.7	93.8	84.2	70.0	100.0	62.0
Claude Opus 4.5 (high)	54.5	71.9	70.6	96.9	71.4	92.0	71.4	93.8	89.5	83.3	97.1	64.8
Grok-4 Fast	47.5	66.7	58.8	90.6	67.9	80.0	66.7	93.8	84.2	70.0	97.1	60.6
Grok-4	48.5	68.4	58.8	93.8	71.4	80.0	71.4	93.8	84.2	70.0	100.0	62.0
Open-Weight Models												
Qwen3-VL-235B-Instruct	38.4	64.9	52.9	90.6	57.1	80.0	57.1	90.0	78.9	60.0	97.1	57.7
Qwen3-VL-235B-Thinking (3-vote)	50.5	68.4	70.6	93.8	82.1	84.0	76.2	93.8	86.8	86.7	100.0	67.6
Qwen3-VL-30B-Thinking	48.5	66.7	61.8	84.4	71.4	72.0	71.4	90.0	89.5	73.3	100.0	57.7
GLM-4.6V	45.5	66.7	58.8	93.8	67.9	76.0	66.7	93.8	92.1	73.3	100.0	64.8
ERNIE-4.5-VL	29.3	58.8	47.1	84.4	46.4	72.0	52.4	83.8	63.2	53.3	94.1	46.5
Mistral-Large	33.3	54.4	47.1	90.6	57.1	68.0	52.4	86.2	68.4	56.7	97.1	43.7
LLaMA-4 Maverick	31.3	50.9	41.2	84.4	50.0	72.0	47.6	83.8	68.4	43.3	94.1	45.1

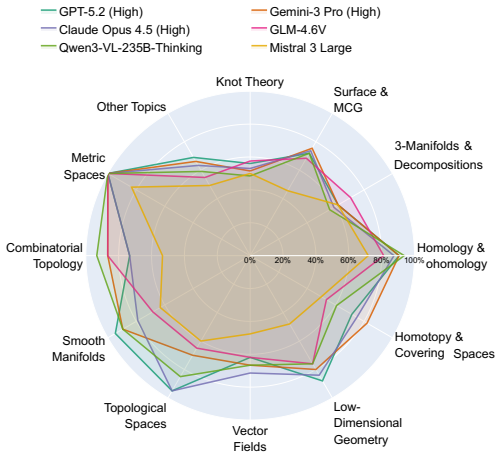


Figure 6: **Core performance across 12 fine-grained topology subfields.** Representative single-run configurations are shown; parentheses indicate inference-time reasoning effort.

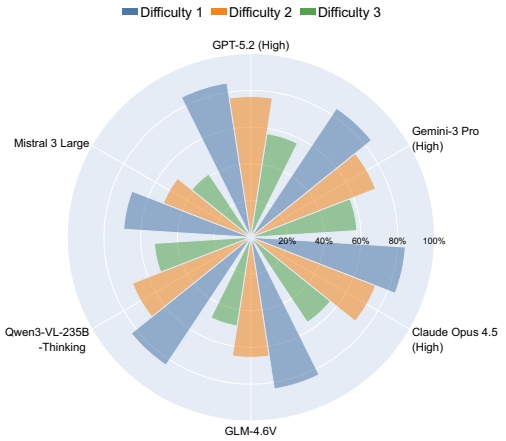


Figure 7: **Core performance across three difficulty levels.** Representative single-run configurations are shown from Level 1 (easiest) to Level 3 (most difficult).

to 66 items. Proprietary models lead on aggregate accuracy, with varying gaps across subfields. The strongest proprietary configuration is GPT-5.2 (high) at 73.0% (Table 9); the strongest open-weight configuration is Qwen3-VL-235B-Thinking with 3-sample voting at 70.0%, compared with 68.5% for its corresponding single run (Table 8). Accuracy varies substantially across fine-grained subfields. Models perform best on *Metric Spaces* and *Homology & Cohomology*, while *Knot Theory* and *3-Manifolds & Decompositions* are the most challenging displayed subfields. The former often reward definition-driven and symbolic reasoning; the latter require models to maintain global geometric relations across a diagram or construction. Aggregate accuracy therefore combines distinct capabilities and can conceal a gap between knowledge-intensive competence and geometric structure tracking. Appendix F illustrates this gap with representative cases.

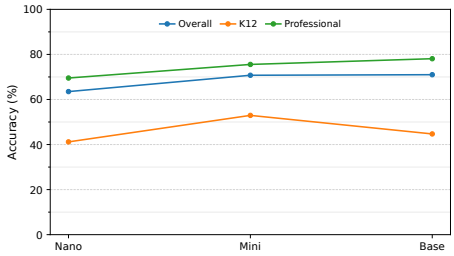
Table 3: **Performance by image dependency on TopoEval-Core.** Dependency columns report accuracy (%) with relative change to overall accuracy in small text; all rows are single-run conditions.

Model	Overall	No image	Present only	Image required
GPT-5 minimal	60.8	61.4 +1.1%	78.9 +30.0%	56.5 -6.9%
GPT-5 high	70.0	84.3 +20.5%	86.0 +22.8%	61.9 -11.5%
GPT-5.2 none	62.0	63.9 +3.0%	75.4 +21.7%	58.5 -5.7%
GPT-5.2 high	73.0	89.2 +22.1%	82.5 +13.0%	65.8 -9.9%
Gemini-3 Flash	70.8	85.5 +20.9%	75.4 +6.6%	65.0 -8.1%
Gemini-3 Pro high	71.8	86.7 +20.9%	86.0 +19.8%	63.8 -11.0%
Claude Opus low	68.2	84.3 +23.6%	84.2 +23.4%	59.6 -12.7%
Claude Opus high	71.0	88.0 +23.9%	84.2 +18.6%	62.7 -11.7%
Grok-4 Fast	64.0	84.3 +31.8%	82.5 +28.8%	53.5 -16.5%
Grok-4	66.5	85.5 +28.6%	89.5 +34.5%	55.4 -16.7%
Qwen3-VL Instruct	58.8	57.8 -1.6%	78.9 +34.4%	54.6 -7.0%
Qwen3-VL Thinking (single-run)	68.5	90.4 +31.9%	75.4 +10.1%	60.0 -12.4%
GLM-4.6V	66.0	75.9 +15.0%	82.5 +24.9%	59.2 -10.3%
ERNIE-4.5-VL	52.8	60.2 +14.2%	71.9 +36.4%	46.2 -12.5%
Mistral-Large	54.0	62.7 +16.0%	71.9 +33.2%	47.3 -12.4%
LLaMA-4 Maverick	52.5	65.1 +23.9%	75.4 +43.7%	43.5 -17.2%

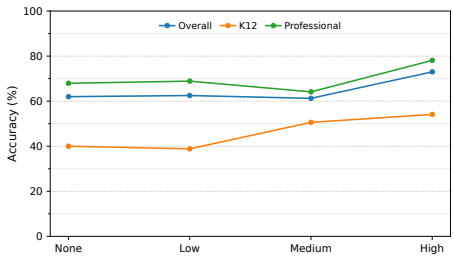
Difficulty levels. Figure 7 compares the three difficulty levels. Across the displayed configurations, accuracy generally decreases from Level 1 to Level 3, supporting the intended ordering by reasoning demand. Variation among models shows that effective difficulty also depends on how well their visual and mathematical representations fit the problem structure.

Image dependency. Table 3 reports Core accuracy for questions with no image, with a non-essential image, and with required visual evidence. Relative changes from overall accuracy range from -1.6% to +31.9%, from +6.6% to +43.7%, and from -17.2% to -5.7%, respectively. The high raw accuracy of the non-essential-image group reflects its answerable textual path and slice composition. The matched control provides a complementary test: removing those images raises accuracy by 1.75–5.26 points across three models. Together, these results show that robust grounding requires both extracting topology-bearing relations from essential images and suppressing incidental cues when text suffices; Appendix B.3 gives the paired results.

K12 vs. Professional. Figure 8 shows a persistent K12–professional gap under parameter and inference-time scaling. Professional-level accuracy improves more reliably, while K12 accuracy remains low and sometimes fluctuates or declines. Within image-required items, macro-averaging the 44 complete Core runs gives 37.1% accuracy for K12 and 62.1% for professional problems, preserving a 24.9-point gap after controlling for image dependency (Table 10). Across the tested scaling conditions, professional knowledge improves more consistently than intuitive structure manipulation. Figure 9 places the 96.3% graduate-student reference beside the strongest complete Core runs, and Appendix C.1 reports the stratified comparison.



(a) GPT-5: parameter-side scaling (Nano → Mini → Base).



(b) GPT-5.2: inference-time scaling (reasoning effort).

Figure 8: **Academic-tier performance on TopoEval-Core: K12 vs. Professional.** Model trends are reported across the Core academic-tier partition.

Proof-track snapshot. Figure 10 reports five strong or representative TopoEval-Proof conditions. Candidate proofs are assessed for answer alignment, use of core definitions or theorems, and logical closure; sage and amber denote strict pass and borderline. Pass-plus-borderline rates remain low. Table 12 further shows that geometric-topology tasks have lower rates than algebraic and point-set/general tasks in the strongest selected condition. The pattern localizes the research-level bottleneck in maintaining definitions, theorem applicability, and logical closure across a sustained argument, motivating process-level supervision in addition to final-answer training. Appendices D–D.2 report the audit protocol and additional results, and Appendix F gives qualitative examples.

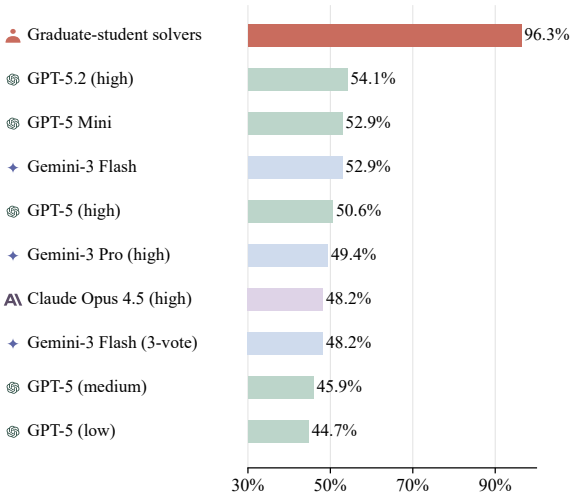


Figure 9: **K12 model–human reference comparison.** Bars use a 30% baseline and report exact K12 accuracy at their endpoints. The graduate-student reference is shown with the strongest complete Core runs.

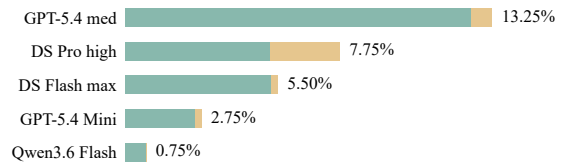


Figure 10: **Selected TopoEval-Proof conditions.** Stacked bars show strict pass (sage) and borderline (amber) on the 400-item Proof track; labels report pass + borderline.

Cross-result synthesis. The same capability boundary appears across these slices. Scaling improves professional and knowledge-intensive items more consistently than K12 or image-required items, and extra inference often preserves the initial visual interpretation. On Proof, locally plausible steps still fail when definitions, theorem conditions, and pending obligations must remain aligned across a long argument. Both tracks therefore point to global state maintenance as a central bottleneck: preserving deformation and connectivity structure in visual problems, and preserving proof obligations through mathematical derivations.

Interpreting progress. These results sharpen what an aggregate gain represents. Core improvements can arise from stronger access to advanced concepts or from better manipulation of visual structure, and academic-tier and image-dependency slices separate these sources. The scaling curves are consistent with additional inference helping most when the initial structural reading is usable. Proof adds a temporal dimension by testing whether goals, assumptions, theorem conditions, and intermediate obligations remain aligned as an argument unfolds. Together, these views provide a diagnostic map of knowledge access, visual grounding, and long-horizon mathematical control.

Prompting versus internal reasoning. Across Qwen3-VL scales, external CoT shifts overall accuracy by -2.3 to $+0.5$ points and voting by -1.5 to $+1.5$ points, whereas thinking variants gain 9.7–11.0 points over instruct counterparts (Table 8). This contrast suggests that repeated decoding alone seldom repairs an incorrect structural representation. Together with academic-tier and image-dependency slices, these variants help distinguish search, knowledge access, and visual interpretation.

5 Conclusion

We introduce TopoEval, an 800-item suite with a 400-item answer-based Core track and a 400-item Proof track. Core reveals persistent errors on K12 and image-required items and a large gap to the graduate-student reference; Proof shows low pass-plus-borderline rates under the reported rubric. The central mismatch is that models perform better on professional, knowledge-intensive Core items than on K12 visual-intuitive ones. This model pattern contrasts with prior evidence of early human sensitivity to connectivity, holes, and continuous deformation (Chien et al., 2012).

The visual results motivate global and deformation-aware representations. Graph-structured or multi-resolution encoders could make connectivity, enclosure, and holes explicit, while evidence-selection mechanisms suppress incidental visual cues. Temporal deformation data, iterative inspection, topology-aware objectives, and homeomorphism-inspired augmentations could further encourage invariance across topologically equivalent shapes.

Research-level proofs expose gaps in target alignment, theorem use, intermediate justification, and logical closure. These tasks require sustained reasoning and command of specialized concepts and recent research contexts. Process supervision over definitions, invariants, theorem applicability, and closure offers a direct training target.

Limitations

Natural-Language Proof Scope. TopoEval-Proof evaluates natural-language proof attempts through reference-grounded rubric judgments, reflecting the explanatory form of research mathematics (Thurston, 1994). The rubric scores target alignment, use of definitions and theorems, and logical closure. Machine-checked evaluation with Lean, Coq, or Isabelle would use formalized statements and checker-specific protocols (Naskręcki and Ono, 2025; Patel et al., 2023; Azerbayev et al., 2023; Poiroux et al., 2025). Core reports answer accuracy, and Proof reports pass, borderline, and fail rates.

Model and API Drift. Several evaluated systems are proprietary or served through external APIs. Their weights, safety layers, routing, and reasoning-effort implementations may change over time, so the reported numbers are snapshots under recorded prompts and settings. Versioned reruns are needed to measure score changes.

Interactive and Agentic Evaluation. The main benchmark comparisons use fixed prompts and call budgets. Figure 22 presents an exploratory five-turn o3 image-inspection trajectory, illustrating how intermediate visual hypotheses can evolve under repeated inspection. A single trajectory does not estimate agentic performance and is therefore excluded from the aggregate comparisons. Systematic evaluation would require standardized tool access, stopping criteria, interaction and compute budgets, and repeated runs to distinguish model capability from trajectory-level variance. Interactive visual methods (Chen et al., 2024; Fu et al., 2025) and tactic-level agents (Lin et al., 2025) remain important extensions, but entail substantially higher and more variable evaluation costs.

Topological Transformation-Based Synthesis. The current suite is expert-curated from educational and academic sources. Future extensions could use homeomorphism-based transformations to generate controlled variants and reduce leakage risk (Ramos et al., 2025); doing so reliably requires computational geometry and topology-aware validation (Hristov et al., 2025).

Ethical Statement

TopoEval is constructed from public educational and academic resources and is intended for non-profit research on foundation model evaluation. We discourage uses that would facilitate academic dishonesty or misrepresent model competence in educational settings.

Human annotators were recruited from project members and research collaborators, and K12 solvers were recruited from topology-trained graduate students. All participated voluntarily without compensation. They were informed of the study context and intended publication and consented to aggregate use of their submitted annotations and answers. Released results aggregate these contributions and exclude participant identifiers and private information.

AI assistants supported code development and language editing. The authors reviewed the resulting code and prose and take responsibility for the final content. Research-pipeline uses of AI systems, including OCR, consistency checks, and evaluation, are described in Sections 3.2–4.1 and the appendices.

During data dissemination, we apply the following copyright and release safeguards:

- **Image Resource Management:** The current release omits image files because the source records do not document row-level redistribution terms. Public records retain the question text, standard answer, metadata, source information, and image references.
- **Content Reconstruction:** To reduce direct reproduction of source materials, questions are substantially rewritten or paraphrased. When appropriate, visual information is converted into equivalent abstract representations while preserving the intended topological structure.

References

Colin C. Adams. 2004. *The Knot Book: An Elementary Introduction to the Mathematical Theory of Knots*. American Mathematical Society.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35*.

Anthropic. 2025a. Introducing claude opus 4.5. Announcement.

Anthropic. 2025b. System card: Claude Opus 4.5. System card. Accessed: 2026-05-26.

Jeremy Avigad. 2018. The mechanization of mathematics. *Notices of the American Mathematical Society*, 65(6):681–690.

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. 2023. ProofNet: Autoformalizing and formally proving undergraduate-level mathematics. In *Advances in Neural Information Processing Systems 36, Datasets and Benchmarks Track*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, and 1 others. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint*.

Baidu ERNIE Team. 2025. ERNIE 4.5 technical report. Technical report, Baidu. Technical report.

Caribou Mathematics Contest. 2025. Caribou contests. Competition website. Accessed: 2026-01-05.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal LLM's referential dialogue magic. *Preprint*, arXiv:2306.15195.

Zhenfang Chen, Qinhong Zhou, Yikang Shen, Yining Hong, Zhiqing Sun, Dan Gutfreund, and Chuang Gan. 2024. Visual chain-of-thought prompting for knowledge-based visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1254–1262.

Zhengyu Chen, Siqi Wang, Teng Xiao, Yudong Wang, Shiqi Chen, Xunliang Cai, Junxian He, and Jingang Wang. 2025. Revisiting scaling laws for language models: The role of data quality and training strategies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23881–23899, Vienna, Austria. Association for Computational Linguistics.

Sarina Hui-Lin Chien, Yun-Lan Lin, Wenli Qian, Ke Zhou, Ming-Kuan Lin, and Hsin-Yueh Hsu. 2012. With or without a hole: Young infants' sensitivity for topological versus geometric property. *Perception*, 41(3):305–318.

Leonardo de Moura, Soonho Kong, Jeremy Avigad, Floris van Doorn, and Jakob von Raumer. 2015. The Lean theorem prover (system description). In *Automated Deduction – CADE-25*, volume 9195 of *Lecture Notes in Computer Science*, pages 378–388. Springer.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Richard Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. 2025. ReFocus: Visual editing as a chain of thought for structured image understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 17783–17805. PMLR.

Demis Hassabis, Koray Kavukcuoglu, and Gemini Team. 2025. A new era of intelligence with Gemini 3. Google Blog. Accessed: 2026-05-26.

Linyang He, Qiyao Yu, Hanze Dong, Baohao Liao, Xinxing Xu, Micah Goldblum, Jiang Bian, and Nima Mesgarani. 2026. LiveMathematicianBench: A live benchmark for mathematician-level reasoning with proof sketches. *arXiv preprint*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint*.

Petar Hristov, Ingrid Hotz, and Talha Bin Masood. 2025. Robust geometric predicates for bivariate computational topology. *arXiv preprint*.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint*.

Sayash Kapoor and Arvind Narayanan. 2023. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*.

John M. Lee. 2012. Simply connected spaces. Course notes, University of Washington.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305. Association for Computational Linguistics.

Haohan Lin, Zhiqing Sun, Sean Welleck, and Yiming Yang. 2025. Lean-STaR: Learning to interleave thinking and proving. In *The Thirteenth International Conference on Learning Representations*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *Preprint*, arXiv:2402.00253.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Wenjie Ma, Andrei Cojocaru, Neel Kolhe, Bradley Louie, Robin Said Sharif, Haihan Zhang, Vincent Zhuang, Matei Zaharia, and Sewon Min. 2025. Reliable fine-grained evaluation of natural language math proofs. *arXiv preprint*.

Mayug Maniparambil, Nils Hoehing, Janak Kapuriya, Arjun Karuvally, Ellen Rushe, Anthony Ventresque, Noel O'Connor, and Fergal Reid. 2026. TopoBench: Benchmarking LLMs on hard topological reasoning. *arXiv preprint*.

Math Kangaroo Association. 2025. Math kangaroo. Competition website. Accessed: 2026-01-05.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. Official announcement. Accessed: 2026-05-26.

Mistral AI. 2025. Introducing mistral 3. Official announcement. Accessed: 2026-05-26.

James R. Munkres. 1984. *Elements of Algebraic Topology*. Addison-Wesley.

James R. Munkres. 2000. *Topology*, second edition edition. Prentice Hall, Inc., Upper Saddle River, NJ.

Bartosz Naskręcki and Ken Ono. 2025. Mathematical discovery in the age of artificial intelligence. *Nature Physics*.

Tobias Nipkow, Lawrence C. Paulson, and Markus Wenzel. 2002. *Isabelle/HOL: A Proof Assistant for Higher-Order Logic*. Springer.

OpenAI. 2025a. Introducing GPT-5.1 for developers. Official announcement. Accessed: 2026-05-26.

OpenAI. 2025b. Update to GPT-5 system card: GPT-5.2. System card. Accessed: 2026-05-26.

OpenRouter. 2024. Openrouter: A unified interface for large language models. Software platform. Accessed: 2025-01.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*.

Nilay Patel, Rahul Saha, and Jeffrey Flanigan. 2023. A new approach towards autoformalization. *arXiv preprint*. ArXiv preprint.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *Preprint*, arXiv:2306.14824.

Auguste Poiroux, Gail Weiss, Viktor Kunčak, and Antoine Bosselut. 2025. Reliable evaluation and benchmarks for statement autoformalization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17947–17969. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. Technical report, Qwen Team.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Patrick Ramos, Ryan Ramos, and Noa Garcia. 2025. Data leakage in visual datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 6368–6378.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint*.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, and Xia Hu. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint*.

William P. Thurston. 1994. On proof and progress in mathematics. *Bulletin of the American Mathematical Society*, 30(2):161–177.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2022. Will we run out of data? limits of LLM scaling based on human-generated data. *arXiv preprint*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with MATH-Vision dataset. In *Advances in Neural Information Processing Systems 37*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint*.

xAI. 2025a. Grok 4. Official announcement. Accessed: 2026-05-26.

xAI. 2025b. Grok 4.1 model card. Model card (PDF). November 17, 2025.

Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P. Lillicrap, Kenji Kawaguchi, and Michael Shieh. 2024. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint*.

Jincheng Xiong. 2011. *Lecture Notes on Point-Set Topology*. Higher Education Press, Beijing. In Chinese.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Chengye You. 1997. *Lecture Notes on Basic Topology*. Peking University Press, Beijing. In Chinese.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Z.ai. 2025. GLM-4.6V. Developer documentation. Accessed: 2026-05-26.

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. 2024. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? In *Computer Vision – ECCV 2024*, volume 15066 of *Lecture Notes in Computer Science*, pages 169–186. Springer.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2022. miniF2F: A cross-system benchmark for formal olympiad-level mathematics. In *Proceedings of the Tenth International Conference on Learning Representations*.

A Benchmark Reference and Positioning

The appendices begin with the terminology, model mappings, topology taxonomy, and benchmark comparisons used throughout the paper. Later sections separate evaluation protocols from supplementary results, followed by reproducibility materials and qualitative cases.

A.1 Mappings and terminology

Table 4: **Parameter-side scaling mappings.** Within each family, `scale` indexes parameter size from smaller to larger.

Family	Scale	Full name
GPT-5	1	GPT-5 (Nano)
	2	GPT-5 (Mini)
	3	GPT-5 (Base)
Qwen3-VL	1	Qwen3-VL-8B
	2	Qwen3-VL-30B
	3	Qwen3-VL-235B
Ministral / Mistral	1	Ministral-3B-2512
	2	Ministral-8B-2512
	3	Mistral-Large-2512
LLaMA-4	1	LLaMA-4 Scout
	2	LLaMA-4 Maverick

Table 5: **General abbreviations and notation used in TopoEval.**

Term	Meaning
Core	TopoEval-Core, the 400-item answer-based track.
Proof	TopoEval-Proof, the 400-item reference-grounded proof-auditing track.
MCQ	Multiple-choice question.
FIB	Fill-in-the-blank question.
K12	Pre-university or school-level topology-intuition item.
L1/L2/L3	Difficulty levels, from easier to harder; Proof levels are empirical tertiles.
CoT	Chain-of-thought prompting.
3-vote	Three-sample majority voting.
P/B	Strict pass count / borderline count in Proof auditing.
P+B	Strict pass plus borderline, a compact Proof-track summary.
LLM	Large language model.
MLLM / LMM / VLM	Multimodal or vision-language foundation model.
OCR	Optical character recognition.
JSON	JavaScript Object Notation; used for structured model outputs.

A.2 Benchmark positioning

Table 6 situates TopoEval by evaluation object, topology coverage, academic span, output assessment, and diagnostic resolution.

Evaluation object. Broad mathematics benchmarks measure general problem solving across subjects, multimodal benchmarks study whether models can use diagrams and other visual evidence, and formal-proof resources target machine-checkable statements or autoformalization. TopoBench (Maniparambil et al., 2026) provides a directly relevant recent comparison: it uses six generated grid-puzzle families to study global spatial constraints such as connectivity, loop closure, region symmetry, reflection, and contiguity. Its controlled instances and rule-based verifiers support detailed analysis of state tracking and constraint propagation. TopoEval instead treats topology as the object of study across heterogeneous problem forms. It asks whether a model can recognize a shared invariant in a visual puzzle, express that invariant mathematically, and use the relevant definitions or theorems when the task advances to professional material. This progression distinguishes disciplinary understanding from success on one controlled puzzle formalism.

Domain depth. TopoEval organizes 800 curated items around topology as a mathematical discipline. Core spans point-set, geometric, algebraic, and differential topology across twelve subfields and K12-to-professional levels, with image-dependency labels for visual-grounding analysis. Its items range from intuitive judgments about holes, connectivity, knots, and deformation to questions requiring explicit invariants, constructions, and theorem applicability. Proof adds reference-grounded assessment of professional natural-language proof attempts, including answer alignment, definition and theorem use, and logical closure. The two tracks therefore follow a common conceptual axis at different depths: Core measures whether the model reaches the right mathematical object or answer, while Proof tests whether it can sustain the argument needed to justify that object at research level.

Community use. This organization supports aggregate model comparison and localized diagnosis by subfield, academic tier, image use, and proof outcome. It separates failures of visual grounding from failures of concept retrieval, theorem selection, and proof closure, allowing a model developer to identify which stage limits a system rather than relying on one overall score. Comparisons

across parameter scale and inference effort further show whether additional capacity improves formal knowledge and intuitive structure tracking at the same rate. More broadly, TopoEval exposes two capability transitions that broad benchmarks tend to blend together: the transition from visual structure recognition to mathematical abstraction, and the transition from concept-level knowledge to a complete research-level argument. Figures 11 and 12 preview the Proof-side source breadth and supplementary model conditions; Appendix D supplies their protocol and interpretation.

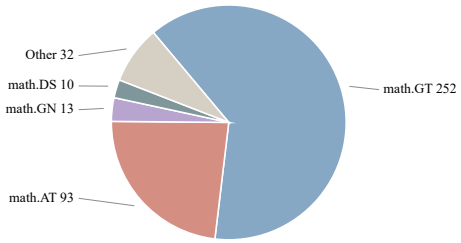


Figure 11: **TopoEval-Proof source-category coverage.** The pie summarizes arXiv source categories using the four largest categories plus an aggregate Other slice.

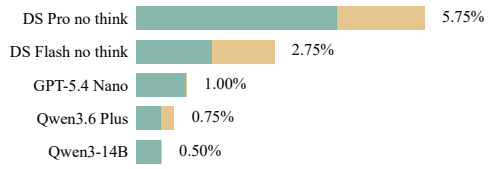


Figure 12: **Additional TopoEval-Proof results.** Stacked bars separate strict pass (sage) and borderline (amber); labels report pass plus borderline on the separate 400-item Proof track. Rates remain low across these additional model and inference conditions.

B Core Evaluation Protocol and Validation

This section documents the Core evaluation interface, the human K12 protocol, and validation checks for answer scoring and image dependency.

B.1 Inference variants and aggregation

Table 8 summarizes the auxiliary prompting, thinking, and voting configurations used to test whether additional inference changes Core accuracy.

Table 8: **Core inference variants and aggregation.** Overall accuracy (%); Δ is the change from the same model's base setting.

(a) Prompt and thinking variants

Model line	Base	Base+CoT	Thinking
Qwen3-VL-8B	48.5	46.2	59.5
Qwen3-VL-30B	54.5	55.0	64.8
Qwen3-VL-235B	58.8	57.5	68.5
GLM-4.6V	66.0	62.0	–

(b) Three-sample majority voting

Model line	Base	+3vote	Δ
Gemini-3 Flash	70.8	70.8	+0.0
Qwen3-VL-235B-Thinking	68.5	70.0	+1.5
GLM-4.6V	66.0	64.5	-1.5
Qwen3-VL-235B-Instruct	58.8	59.0	+0.3

B.2 Human and automated evaluation protocols

Human K12 reference. The human reference point covers the audited K12 subset. Solvers were recruited directly from topology-trained graduate students and volunteered without compensation. They received a common protocol: work independently and submit answers through a structured interface within two hours, with access to reference materials. Aggregate correctness is 96.3%. Figure 9 compares this reference with strong Core configurations evaluated on all 85 K12 items. Standardized human protocols for professional Core and Proof remain future extensions.

Automated answer judging. For TopoEval-Core, GPT-5-nano performs structured final-answer extraction, normalization, and consistency checking after a model has produced an answer. The Core tasks are multiple-choice or fill-in-the-blank, so the final correctness decision is made by matching the normalized extracted answer against the gold answer, with straightforward equivalence handling for mathematical symbols and formulas. The answering prompts and judge prompts require strict JSON output, which reduces parsing ambiguity. As described in Section 4.1, manual inspection of 100 automatically judged Core items found one extraction error. Proof uses rubric-level judgments. A complementary review of 100 Core

Table 6: **Evaluation-design comparison with representative mathematics, multimodal, proof, and topology-focused benchmarks.** The columns expose complementary design choices rather than ranking datasets by size. TopoBench studies global constraints in generated grid puzzles; TopoEval provides discipline-wide mathematical topology coverage from visual intuition to professional proof.

Benchmark	Evaluation object	Topology coverage	Visual and level structure	Output assessment	Diagnostic resolution
MATH (Hendrycks et al., 2021)	Competition mathematics problems	Broad mathematics; no dedicated topology taxonomy	Text; competition level	Final answers and worked solutions	Subject and difficulty categories
MathVista (Lu et al., 2024)	Mathematics in visual contexts	Broad mathematics; topology appears within a multi-domain collection	Images and text; mixed skills	Answer-based multimodal reasoning	Task and reasoning-skill categories
MathVerse (Zhang et al., 2024)	Multi-subject visual mathematics	Broad mathematics; no dedicated topology taxonomy	Six modality variants	Answers and intermediate CoT scoring	Visual-information use and reasoning steps
miniF2F (Zheng et al., 2022)	Formal theorem statements	Broad olympiad mathematics	Text; olympiad to early undergraduate	Machine-checked formal proofs	Theorem and proof-assistant setting
ProofNet (Azerbayev et al., 2023)	Natural-language and Lean theorem pairs	Undergraduate mathematics, including topology	Text; undergraduate	Autoformalization and formal proving	Domain and theorem
TopoBench (Maniparambil et al., 2026)	Generated grid puzzles	Six global-constraint families: connectivity, closure, symmetry, reflection, contiguity	ASCII grids and format interventions; three generator tiers	Rule-based verification of complete grids	Puzzle family, tier, trace errors, and causal interventions
TopoEval	**Curated mathematical-topology problems**	**Four branches and twelve fine-grained subfields**	**Text and images; K12 to professional; image-dependency labels**	**Core answer accuracy and reference-grounded natural-language proof auditing**	**Subfield, difficulty, academic tier, image use, and proof outcome**

CoT/response records checked for reliance on option wording or position, unsupported elimination, and guessing; no systematic pattern appeared in this sample. Core also includes 76 fill-in-the-blank items without answer-option cues.

Answer-format adaptation. Core uses 324 multiple-choice and 76 fill-in-the-blank items as a reproducible final-answer interface. Annotators preserved the mathematical substance, verified final answers, and randomly assigned MCQ answer positions. Proof and the qualitative case studies provide proof-level and reasoning-level evidence alongside Core accuracy. Future extensions can pair original and adapted forms under expert-reviewed open-ended criteria, quantify distractor quality, and evaluate stability under option reshuffling.

B.3 Validation checks

Image-dependency categories. The paper-facing visual axis measures image dependency: whether an item can be solved from text and whether visual evidence is required. We distinguish questions with no image, questions where an image is present but not logically required, and questions where the image is required for solving. The middle category tests model behavior when a visual cue is available but unnecessary for solving the item, while the image-required category concentrates tasks that depend on visual evidence. Together, these categories separate sensitivity to incidental images from performance on visually grounded reasoning.

Paired control for non-essential images. To probe the middle dependency category more directly, we evaluated the same 57 image-present-but-not-required (Dep-1) questions with and without their non-essential images under matched answer protocols. Table 9 reports dedicated paired-control runs for three models. These scores form within-row comparisons and use runs distinct from the full-Core results in Table 3. All three score lower with the images present, by 1.75–5.26 percentage points. This matched-item pattern records sensitivity to non-essential visual input under the tested protocols.

Table 9 reports Wilson 95% confidence intervals for selected benchmark proportions, including K12 and image-required subsets.

C Core Supplementary Results

This section extends the main Core analysis with image-controlled K12 results, inference variants, aggregation, and additional subfield rows.

C.1 K12 gap under image-dependency control

K12 items are more often image-required than professional items, so Table 10 stratifies the K12–

Table 7: **Abbreviation mapping for the 12 fine-grained topology subfields.** We summarize each subfield's full name, major branch, and typical scope.

Abbr.	Full name	Scope (major branch + brief description)
TopSp	Topological Spaces	**Point-set.** Core axioms and basic notions of topological spaces, including open and closed sets, bases, and continuity.
Metric	Metric Spaces	**Point-set.** Metric structures, convergence, completeness, and relationships between metric and topological notions.
Knot	Knot Theory	**Geometric.** Knots and links, isotopy, invariants, and diagrammatic reasoning about entanglement.
Surf	Surface Topology & Mapping Class Groups	**Geometric.** Classification and structures of surfaces, mapping class groups, and basic surface invariants.
3M	3-Manifolds & Decompositions	**Geometric.** 3-manifold intuition and standard decompositions, including splitting or cutting along surfaces.
LDG	Low-Dimensional Geometry	**Geometric.** Geometric structures and intuition in low dimensions, including planar and three-dimensional geometric-topological interactions.
Comb	Combinatorial Topology	**Geometric.** Piecewise-linear and combinatorial viewpoints, including simplicial complexes, triangulations, and discrete invariants.
Hom	Homology & Cohomology	**Algebraic.** Homology and cohomology as deformation-invariant descriptors of holes, cycles, and related computations.
Hty	Homotopy & Covering Spaces	**Algebraic.** Homotopy equivalence, fundamental group intuition, covering spaces, and lifting properties.
SmMfd	Smooth Manifolds	**Differential.** Smooth manifolds and smooth maps; local-to-global structure under differentiability constraints.
VFld	Vector Fields	**Differential.** Vector fields, indices, and qualitative behavior on manifolds, including flows and critical behavior.
Other	Other Topics	**Mixed.** Rarer or overlapping subfields, including compactness and connectedness, continuous maps and homeomorphisms, and topological constructions.

(a) Paired Dep-1 image control.

Model	With	Without	Δ
GPT-5.2	78.95	82.46	-3.51
Gemini-3-Flash	71.93	77.19	-5.26
Qwen3-VL-235B-Thinking	80.70	82.46	-1.75

(b) Wilson 95% intervals for selected rates.

Rate	Count	%	95% CI
Core overall, GPT-5.2 high	292/400	73.0	[68.4, 77.1]
Core K12, GPT-5.2 high	46/85	54.1	[43.6, 64.3]
Core image-required, GPT-5.2 high	171/260	65.8	[59.8, 71.3]
Proof P+B, GPT-5.4 medium	53/400	13.3	[10.3, 16.9]

Table 9: **Selected validation summaries.** (a) The same 57 questions are evaluated with and without non-essential images; Δ is with-image minus without-image accuracy in percentage points. (b) P+B denotes pass plus borderline on the Proof track.

professional comparison by image dependency. The no-image and image-present-but-not-required K12 slices are small and serve as transparency checks. In the image-required stratum, which contains 87.06% of K12 items, the 44 complete Core runs show a 24.9 percentage-point professional-over-K12 accuracy gap on average; GPT-5.2 (high) shows a 20.2-point gap. The image-required stratum retains a substantial K12–professional gap, pointing to additional sources of difficulty within visually dependent items.

C.2 Inference variants and aggregation

Table 8 compares external CoT prompting, dedicated thinking variants, and 3-sample voting. External CoT and voting produce small and uneven changes, whereas the dedicated Qwen thinking variants improve consistently over their instruct counterparts. This pattern is consistent with extra sampled text helping once a model has formed a useful representation, while stronger inference behavior inside the model is more reliable when a problem requires tracking a diagram, connectivity relation, or invariant.

C.3 Additional subfield results

Table 11 expands the subfield comparison beyond the representative models in the main paper. Performance remains strongest on definition- and computation-oriented areas such as Metric Spaces and Homology, while several geometric subfields vary sharply across model families. The broader rows reinforce that aggregate accuracy can conceal different balances of formal knowledge and geometric structure tracking.

D TopoEval-Proof Protocol and Supplementary Results

This section records the TopoEval-Proof construction and audit protocol, followed by Proof-specific supplementary results. The track contains 400 professional proof tasks evaluated with pass, borderline, and fail verdicts. Public records include proof statements, local context, taxonomy metadata, and task instructions. Internal audit records add reference proofs and proof obligations for reference-grounded grading. Figure 11 summarizes the source-category distribution: most items come from geometric and algebraic topology cat-

Table 10: **K12 gap under image-dependency control.** Mean columns macro-average 44 complete Core runs with the recorded 85-item K12 and 315-item professional split. Positive gaps indicate higher professional accuracy than K12 accuracy within the same image-dependency stratum; small K12 strata are included for transparency.

Image dependency	K12 n	Prof. n	Complete-run mean			GPT-5.2 (high)		
			K12	Prof.	Gap	K12	Prof.	Gap
No image	6	77	75.0	71.3	-3.7	83.3	89.6	6.3
Present, not required	5	52	48.6	80.9	32.2	60.0	84.6	24.6
Image required	74	186	37.1	62.1	24.9	51.4	71.5	20.2

Table 11: **Additional results on the 12 fine-grained topology subfields.** We report accuracy (%) for additional models beyond Table 2. Red indicates the best performance and Blue indicates the second best within this table (ties for best are all marked in red and no blue is assigned for that column). The parentheses in the model names indicate the level of inference-time reasoning effort.

Model	Overall	Knot	Surf	3M	Hom	Hty	LDG	VFld	TopSp	SmMfd	Comb	Metric	Other
Proprietary Models													
GPT-5-Nano	63.5	45.5	61.4	52.9	75.0	60.7	92.0	52.4	75.0	78.9	80.0	100.0	59.2
GPT-5-Mini	70.8	57.6	68.4	61.8	84.4	71.4	84.0	57.1	95.0	94.7	73.3	100.0	63.4
GPT-5 (low)	66.8	47.5	66.7	55.9	81.2	75.0	80.0	66.7	80.0	86.8	70.0	100.0	60.6
GPT-5 (medium)	68.5	48.5	71.9	61.8	90.6	71.4	84.0	61.9	80.0	84.2	80.0	100.0	62.0
GPT-5.2 (low)	65.5	46.5	68.4	55.9	84.4	67.9	76.0	57.1	75.0	84.2	70.0	100.0	60.6
GPT-5.2 (medium)	61.3	53.0	23.0	23.0	87.5	42.9	84.0	66.7	80.0	84.2	73.3	100.0	49.4
Gemini-3 Pro (low)	64.8	45.5	61.4	61.8	90.6	71.4	76.0	66.7	85.0	84.2	70.0	100.0	52.1
Claude Haiku 4.5	58.8	41.4	63.2	55.9	84.4	60.7	72.0	52.4	80.0	78.9	56.7	97.1	52.1
Claude Opus 4.5 (medium)	68.0	47.5	66.7	58.8	90.6	75.0	84.0	66.7	90.0	89.5	70.0	100.0	60.6
Grok-4.1 Fast	65.2	46.5	66.7	58.8	90.6	67.9	80.0	66.7	93.8	84.2	70.0	100.0	56.3
Open-Weight Models													
Qwen3-VL-8B-Instruct	48.5	40.9	43.9	47.1	43.8	71.4	52.0	57.1	50.0	42.1	40.0	66.7	49.3
Qwen3-VL-8B-Thinking	59.5	48.5	57.9	58.8	75.0	60.7	64.0	42.9	70.0	63.2	73.3	100.0	53.5
Qwen3-VL-30B-Instruct	56.5	44.4	61.4	52.9	87.5	60.7	76.0	61.9	90.0	68.4	56.7	97.1	50.7
Qwen3-VL-30B-Thinking	65.0	48.5	66.7	61.8	84.4	71.4	72.0	71.4	90.0	89.5	73.3	100.0	57.7
LLaMA-4 Scout	45.2	28.3	50.9	35.3	75.0	39.3	64.0	42.9	80.0	57.9	40.0	80.0	42.3
Ministral-3B-2512	33.8	18.2	33.3	29.4	53.1	25.0	52.0	28.6	68.8	42.1	20.0	52.9	31.0
Ministral-8B-2512	44.2	25.3	47.4	35.3	75.0	39.3	68.0	42.9	83.8	63.2	30.0	76.5	36.6
Ministral-14B-2512	46.5	29.3	50.9	35.3	78.1	39.3	64.0	47.6	86.2	63.2	36.7	80.0	36.6
Mistral-Large-2512	51.8	31.3	52.6	41.2	90.6	50.0	68.0	52.4	83.8	73.7	53.3	97.1	40.8

egories, while general, differential, and smaller adjacent categories provide additional breadth.

D.1 Dataset and audit protocol

Audit rubric. Each candidate proof is graded by a reference-aware auditor on three 0–2 dimensions: *answer alignment*, whether the proof addresses the exact statement; *core definition/theorem usage*, whether it invokes the necessary topology definitions, constructions, or theorems; and *logical closure*, whether the argument forms a valid chain from assumptions to conclusion. A proof is marked *pass* when all three dimensions score 2 and no critical flaw is present. It is marked *borderline* when no dimension scores 0 and at least two dimensions score 2. It is marked *fail* if any dimension scores 0 or if the judge identifies a critical flaw. We report strict pass and borderline separately; pass-plus-borderline is the combined summary under the Proof rubric.

Audit check and scope. Manual inspection of 30 Proof records covered candidate excerpts, verdicts, and dimension-level rationales, and found decisions broadly consistent with the reference-grounded rubric. This review serves as a sanity check for natural-language proof quality under the rubric; future work can quantify human and independent second-judge agreement (Ma et al., 2025).

D.2 Additional Proof-track results

Appendix C reports the supplementary Core analyses for the multiple-choice and fill-in-the-blank

items under answer accuracy. We keep the Proof results separate because their evaluation unit is a complete natural-language argument and their outcomes are rubric-based verdicts rather than final-answer correctness. Table 12 breaks selected Proof conditions down by high-level topology branch, reporting strict pass (P), borderline (B), and their combined P+B rate. Figure 12 complements the main Proof snapshot with additional model and inference conditions at the aggregate level.

Table 12: **TopoEval-Proof selected results by high-level subfield.** Each cell reports P/B followed by the P+B rate within that subfield. Subfield counts are Geometric 233, Algebraic 100, Point-set/General 38, and Differential 29.

Model	Setting	Geometric	Algebraic	Point-set/General	Differential
GPT-5.4	medium	19/2; 9.0%	18/1; 19.0%	9/0; 23.7%	4/0; 13.8%
DeepSeek v4 Pro	high thinking	7/5; 5.2%	8/5; 13.0%	3/0; 7.9%	3/0; 10.3%
DeepSeek v4 Pro	no thinking	6/4; 4.3%	7/2; 9.0%	3/1; 10.5%	0/0; 0.0%
DeepSeek v4 Flash	max thinking	4/1; 2.1%	11/0; 11.0%	5/0; 13.2%	1/0; 3.4%
DeepSeek v4 Flash	no thinking	2/2; 1.7%	3/0; 3.0%	0/3; 7.9%	1/0; 3.4%
GPT-5.4 Mini	default	3/0; 1.3%	3/1; 4.0%	3/0; 7.9%	1/0; 3.4%
GPT-5.4 Nano	default	1/0; 0.4%	2/0; 2.0%	0/0; 0.0%	1/0; 3.4%
Qwen3.6 Flash	no thinking	2/0; 0.9%	0/0; 0.0%	1/0; 2.6%	0/0; 0.0%
Qwen3.6 Plus	no thinking	0/0; 0.0%	1/0; 1.0%	1/1; 5.3%	0/0; 0.0%
Qwen3-14B	no thinking	1/0; 0.4%	0/0; 0.0%	1/0; 2.6%	0/0; 0.0%

Interpretation. Across the selected conditions in Table 12, P+B rates remain low even when models receive larger inference budgets. GPT-5.4 medium is the strongest selected configuration: its P+B rate reaches 23.7% on point-set/general topology and 19.0% on algebraic topology, compared with 13.8% on differential and 9.0% on geometric topology. Because geometric topology contributes 233 of the 400 Proof items, its lower rates on geometric constructions, spatial relations, and low-dimensional objects weigh heavily on the aggregate score.

DeepSeek v4 Pro with high thinking is the strongest selected non-GPT condition, but its gains over the no-thinking setting are not uniform across branches. P+B rises from 4.3% to 5.2% on geometric topology, from 9.0% to 13.0% on algebraic topology, and from 0.0% to 10.3% on differential topology, while point-set/general performance falls from 10.5% to 7.9%. Figure 12 shows the same broader pattern: the additional no-thinking and smaller-model conditions remain below 6% P+B. More inference can therefore recover some specialized knowledge or missing steps, but it does not reliably produce the target alignment, theorem use, and logical closure required for a complete research-level proof.

E Reproducibility Materials

E.1 Public release and internal audit boundary

The public package is available at `https://github.com/YinzhuCheng/TopoEval`. It supports record inspection, Core scoring, Proof generation, and reproduction of aggregate analyses through released Core rows, answers, metadata, scoring code, Proof candidate-facing fields and prompts, rubric schemas, tables, and validation summaries. Core records retain image references; files lacking documented redistribution terms are withheld. Reference-bearing Proof fields, item-specific audit payloads, hidden audit subsets, and provider call records remain controlled audit materials.

Contamination risk and source recency. Core items are adapted from public textbooks, lecture notes, competitions, and evaluation systems, leaving possible training-corpus overlap. Rewriting, answer-format adaptation, source traceability, and human verification reduce exact textual overlap. All 400 Proof items in the current snapshot come from arXiv papers posted December 16, 2025–May 21, 2026; 365 postdate January 1 and 141 postdate April 1. Following live mathematics benchmarks built from newly published papers (He et al., 2026), we release source metadata and recommend versioned reruns as exposure changes.

E.2 Prompting protocols

Six displayed templates cover Core answer generation and Proof generation and auditing. Core crosses multiple-choice versus free-form output with CoT versus answer-only generation, while a separate judge instruction normalizes answers for gold matching. Proof separates candidate generation from reference-aware auditing; the auditor scores alignment, definition or theorem use, and closure on a 0–2 scale and returns structured verdict fields.

E.2.1 Core answering prompts

The four templates form two matched pairs: A and B require normalized option letters, while C and D accept typed scalar or list-valued answers. Within each pair, only the explicit reasoning request and corresponding JSON field change.

Answering Model Prompt A (MCQ, CoT=on)

```
You will answer a multiple-choice question. Some questions may have multiple correct options.
You SHOULD output your chain-of-thought reasoning.
Output format rules (STRICT):
- Return ONLY a JSON object. No markdown. No extra text.
- JSON schema:{ "answer": "A" | "A,B,C", "cot": string }
- "answer" must contain ONLY letters among A,B,C,D,E.
- If multiple, separate by comma "," with NO spaces (example: "A,B,C").
- "cot" is your step-by-step reasoning.
Example:{ "answer": "B", "cot": "..." }
Question:<QUESTION>
Options:<OPTIONS>
```

Answering Model Prompt B (MCQ, CoT=off)

```
You will answer a multiple-choice question. Some questions may have multiple correct options.
Output format rules (STRICT):
- Return ONLY a JSON object. No markdown. No extra text.
- JSON schema:{ "answer": "A" | "A,B,C" }
- "answer" must contain ONLY letters among A,B,C,D,E.
- If multiple, separate by comma "," with NO spaces (example: "A,B,C").
Example:{ "answer": "B" }
Question:<QUESTION>
Options:<OPTIONS>
```

Answering Model Prompt C (Free-form, CoT=on)

```
Answer the following question.
You SHOULD output your chain-of-thought reasoning.
Output format rules (STRICT):
- Return ONLY a JSON object. No markdown. No extra text.
- JSON schema:{ "answer": string | number | boolean | list[string], "cot": string }
- "answer" must be the final result.
- For fill-in with multiple blanks, you may output a list of strings.
- "cot" is your step-by-step reasoning.
Question:<QUESTION>
```

Answering Model Prompt D (Free-form, CoT=off)

```
Answer the following question without analysis.
Output format rules (STRICT):
- Return ONLY a JSON object. No markdown. No extra text.
- JSON schema:{ "answer": string | number | boolean | list[string] }
- "answer" must be the final result only (concise, no derivations).
Question:<QUESTION>
```

E.2.2 Proof generation and judging prompts

The generator sees only candidate-facing information and produces a self-contained natural-language proof. The judge is invoked afterward with private reference material and item-specific obligations. The templates reproduce the implemented message structures while replacing private content with generic field markers.

Proof-track Candidate Generation Prompt

```
System:
You are an advanced topology reasoning model.
Write a complete natural-language proof for the given topology problem.
Do not cite external papers, theorem numbers, or hidden sources.
If a key step is uncertain, state the claim and justify it as far as possible.
Return only the proof text.

User JSON:
{
  "problem_statement": "<PUBLIC_PROOF_STATEMENT>",
  "instructions": "<ITEM_INSTRUCTIONS>",
  "local_context": {
    "allowed_lemmas": ["..."], "definitions": ["..."], "notation": ["..."]
  },
  "answer_format": "free_form_proof",
  "important": [
    "Do not cite external papers, theorem numbers, labels, or hidden sources.",
    "Use only the public statement and local context.",
    "Write a coherent proof, not a list of guesses."
  ]
}
```

Proof-track Reference-aware Judge Prompt

```
System:
You are a reference-aware informal topology proof auditor.
You are not a formal proof verifier.
Grade the candidate proof using the fixed three-dimensional TopoEval-Proof rubric.
Return only valid JSON.

User JSON:
{
  "positioning": "Reference-grounded informal proof auditing, not formal verification.",
  "problem_statement": "<PUBLIC_PROOF_STATEMENT>",
  "candidate_proof": "<CANDIDATE_PROOF>",
  "reference_proof_excerpt": "<PRIVATE_REFERENCE_EXCERPT>",
  "proof_obligations": {
    "required_steps": ["..."], "core_ideas": ["..."],
    "minimal_valid_proof_criteria": ["..."], "common_failure_modes": ["..."]
  },
  "official_3d_rubric": {
    "dimensions": [{"name": "<DIMENSION>", "score_0": "...", "score_1": "...", "score_2": "..."}],
    "allowed_assumptions": ["..."], "disallowed_shortcuts": ["..."],
    "verdict_rule": {"pass": "...", "borderline": "...", "fail": "..."}
  },
  "scoring_rules": {
    "dimensions": ["answer_alignment", "core_definition_theorem_usage", "logical_closure"],
    "score_values": [0,1,2],
    "pass": "all dimensions are 2 and no critical flaw",
    "borderline": "no zero and at least two dimensions are 2",
    "fail": "any zero or a critical flaw"
  },
  "return_schema": {
    "dimension_scores": {"answer_alignment": "0|1|2",
      "core_definition_theorem_usage": "0|1|2", "logical_closure": "0|1|2"},
    "official_verdict": "pass|borderline|fail", "critical_flaw": "boolean",
    "dimension_rationales": {"dimension_name": "short rationale"},
    "covered_reference_points": ["..."], "missing_reference_points": ["..."],
    "unsupported_assumptions": ["..."], "error_tags": ["snake_case_tag"],
    "short_comments_zh": "concise audit summary", "audit_confidence": "0..1"
  }
}
```

F Case Studies

This section pairs four Proof audits with Core cases spanning final answers, reasoning traces, and visual interpretation. We use *knowledge deficiency* for missing or misapplied topology. *Reasoning error* marks inconsistent intermediate steps, while *image understanding error* marks lost visual structure. Together, the cases expose the knowledge–representation gap across textual proofs, diagrams, and deformation tasks.

Proof Case 1: Surface-link quandle isomorphism

Public statement. Let S be a surface link and let $\mathcal{D}$ be a banded unlink diagram. Then $Q(S)$ and $Q(\mathcal{D})$ are isomorphic.

Candidate model and status. GPT-5.4 medium; strict pass.

Candidate proof excerpt. The proof places S in hyperbolic form in $\mathbb{R}^3 \times \mathbb{R}$. At a level below the saddle points, the cross-section is the unlink L; each saddle is represented by a band in $\mathcal{D} = (L, B)$; and the upper and lower components are capped by disks. The candidate then compares the meridional-path quandle of the exterior of S with the diagrammatic presentation from L, crossing relations, and band relations, arguing that the caps add no new quandle relations.

Rubric focus.

Core objects	Surface-link fundamental quandle; banded unlink diagram; meridional paths in the exterior.
Expected route	Put the surface in hyperbolic form, read the lower link quandle from the unlink diagram, add exactly the saddle-band relations, and check that the cap disks do not introduce extra quandle relations.
Strict-pass criterion	The proof must identify the same generators and relations on both sides and tie the isomorphism to the level-by-level construction.

Reference-aware Audit

Verdict. Strict pass.

Dimension-level audit.

Answer alignment	$2/2$	The proof states and proves the intended isomorphism $Q(S) \cong Q(\mathcal{D})$.
Core definition/theorem usage	$2/2$	It grounds the isomorphism in the quandle presentation induced by meridional paths, link crossings, and band moves.
Logical closure	$2/2$	The lower-link, saddle, and cap stages are connected into a complete level-by-level comparison.

Audit rationale. The proof targets the quandle isomorphism $Q(S) \cong Q(\mathcal{D})$, uses the banded unlink presentation, and tracks how crossing and band relations arise through the level decomposition. The level-by-level comparison closes the argument from the unlink stage through the saddle moves.

What this illustrates. The auditor rewards a proof that aligns the model's high-level topological description with the concrete presentation required by the problem.

Figure 13: **Proof-track case study 1.** A strict pass from GPT-5.4 medium. The candidate connects the surface-link invariant to the banded-diagram presentation through explicit generators and relations.

Proof Case 2: Conormal knot triviality

Public statement. Suppose that K_1 has the same knot type as K_0 in $\mathbb{R}^3$. Then $\varphi^{-1}(K_1)$ is a trivial knot in $T^*_{K_0}\mathbb{R}^3$.

Candidate model and status. DeepSeek v4 Pro, high thinking; borderline.

Candidate proof excerpt. The candidate defines $T^*_K\mathbb{R}^3 = \{(q,p) : q \in K,\ p(T_qK) = 0\}$ and treats it as a rank-2 bundle over S^1. It argues that this conormal bundle is trivial, so the zero section is the core $S^1 \times \{0\}$. It then uses an ambient diffeomorphism carrying K_0 to K_1, lifts it to the cotangent bundle, and concludes that the preimage of K_1 is carried to the zero section, hence is unknotted in the conormal bundle.

Rubric focus.

Core objects	Conormal bundle $T^*_{K_0}\mathbb{R}^3$, zero section, lifted ambient diffeomorphism, and trivial knot in the conormal total space.
Expected route	Relate equal knot type to an ambient isotopy, lift the induced map to cotangent data, and justify why the preimage is isotopic to the core in the conormal bundle.
Borderline risk	A proof can have the right geometric picture while still omitting local structural ingredients expected by the reference-aware rubric, such as incompressibility or group-theoretic controls used to rule out hidden twisting.

Reference-aware Audit

Verdict. Borderline.

Dimension-level audit.

Answer alignment	$2/2$	The candidate proves the stated target, namely that $\varphi^{-1}(K_1)$ is trivial in $T^*_{K_0}\mathbb{R}^3$.
Core definition/theorem usage	$1/2$	It correctly invokes conormal bundles and lifted diffeomorphisms; the local theorem usage expected by the reference route is missing.
Logical closure	$2/2$	Within its chosen geometric sketch, the steps form a coherent chain from isotopy to triviality.

Audit rationale. The proof reaches the intended conclusion and gives a coherent geometric route through conormal bundles and lifted diffeomorphisms. The audit marks the core-theorem dimension as incomplete because the candidate omits the local group-theoretic and incompressibility ingredients expected by the reference-grounded proof route. The resulting sketch leaves required structure unestablished.

What this illustrates. Borderline proofs can be globally coherent while still omitting the specific local theorem usage needed for a strict proof audit.

Figure 14: **Proof-track case study 2.** A borderline DeepSeek v4 Pro proof. The conclusion and reasoning chain align with the task; the reference-aware audit records missing core local ingredients.

Proof Case 3: Branched covers of bridge quadrisections

Public statement. Consider a bridge quadrisection

$$(S^5, Y) = (W_1, E_1) \cup (W_2, E_2) \cup (W_3, E_3) \cup (W_4, E_4)$$

and a branched covering $f : Z \to S^5$ along Y, where Z is a connected orientable closed 5-manifold. Then

$$Z = f^{-1}(W_1) \cup f^{-1}(W_2) \cup f^{-1}(W_3) \cup f^{-1}(W_4)$$

is a quadrisection of Z.

Candidate model and status. DeepSeek v4 Flash, max thinking; fail.

Candidate proof excerpt. The candidate defines $Z_i = f^{-1}(W_i)$ and verifies the formal identities for double and triple intersections. It then claims that the branched cover of each sector and each intersection sector is again the required handlebody, because the branch set is locally trivial.

Rubric focus.

Core objects	Bridge quadrisection sectors W_i, preimage sectors $f^{-1}(W_i)$, branched-cover local models, and the handlebody axioms for all intersections.
Expected route	Prove the set-theoretic decomposition of Z and establish the required handlebody type for every lifted sector and every pairwise or triple intersection.
Critical checks	The proof must justify the topology of each W_i, together with the connectedness and handlebody structure of preimages under branching.

Reference-aware Audit

Verdict. Fail.

Dimension-level audit.

Answer alignment	$2/2$	The desired decomposition is explicitly stated.
Core definition/theorem usage	$0/2$	The candidate misapplies a local branched-cover model and treats sector handlebodies as if the required global hypotheses were automatic.
Logical closure	$0/2$	The proof never establishes the quadrisection axioms for the lifted pieces, and the connectedness claim is unsupported.

Audit rationale. The candidate states the desired conclusion, so answer alignment is high. The failure is structural: it applies a local branched-cover model as if each W_i were a ball, without proving the decomposition needed to recover the handlebody structure. It also assumes connectedness of preimages under branching without justification. The central topology argument supporting the stated endpoint is therefore incomplete.

What this illustrates. Proof-track auditing separates a correct-looking conclusion from a valid proof. Here, the quadrisection conclusion additionally requires handlebody and connectedness arguments.

Figure 15: **Proof-track case study 3.** A DeepSeek v4 Flash failure. The model identifies the intended decomposition; its use of the local theorem fails to establish the required handlebody types.

Proof Case 4: Arc-complex distance after puncturing

Public statement. Let S be a connected orientable surface with $b \geq 2$, let $P \subset \mathrm{int}(S)$ be finite, and let $\varphi \in \mathrm{Homeo}(S)$ satisfy $\varphi(P) = P$. For any vertex γ of $\mathcal{A}^\dagger(S)$ disjoint from P,

$$d_{\mathcal{A}^s(S\setminus P)}\left([\gamma]_{S\setminus P}, [\varphi]^i_{S\setminus P}[\gamma]\right) \leq d_{\mathcal{A}^\dagger(S)}(\gamma, \varphi^i(\gamma))$$

for all $i \in \mathbb{Z}$.

Candidate model and status. Qwen3-14B, no thinking; strict pass.

Candidate proof excerpt. The candidate interprets the two arc complexes and their graph metrics, observes that γ and $\varphi^i(\gamma)$ remain disjoint from P, and maps a path in $\mathcal{A}^\dagger(S)$ to the punctured surface by viewing the same arcs in $S \setminus P$. Since disjointness is preserved under this inclusion, the image path has length no greater than the original path.

Rubric focus.

Core objects	Arc complexes $\mathcal{A}^\dagger(S)$ and $\mathcal{A}^s(S \setminus P)$, graph distance, induced isotopy classes after puncturing, and the action of φ.
Expected route	Take a shortest or arbitrary path in the unpunctured arc complex, view each arc in the punctured surface, and use preserved disjointness to obtain a path of no greater length.
Strict-pass criterion	The proof must account for the punctures and induced classes $[\gamma]_{S\setminus P}$, and explicitly establish the graph-metric inequality.

Reference-aware Audit

Verdict. Strict pass.

Dimension-level audit.

Answer alignment	$2/2$	The candidate proves the exact distance inequality in the public statement.
Core definition/theorem usage	$2/2$	It uses the definitions of vertices, disjointness edges, induced punctured-surface classes, and graph distance.
Logical closure	$2/2$	Every edge in the original path remains admissible after puncturing, so the distance upper bound follows.

Audit rationale. The proof directly proves the required distance inequality. It uses the arc-complex definitions, the induced classes in $S \setminus P$, and the fact that preserving a path gives an upper bound on graph distance. The argument closes because every edge of the original path remains an admissible disjointness relation after passing to the punctured surface.

What this illustrates. This case illustrates a short proof route based on a stable functorial or monotonicity argument.

Figure 16: **Proof-track case study 4.** A strict pass from Qwen3-14B. The proof succeeds by correctly transporting an arc-complex path to the punctured surface and reading off the distance inequality.

Problem in Differential Topology

Let M be a smooth, closed, compact, connected, orientable surface in three-dimensional Euclidean space $\mathbb{R}^3$. Denote by K the Gaussian curvature of M, and by $d\sigma$ the area element on M. Consider the total curvature

$$I = \int_M K \, d\sigma.$$

Which of the following values can I attain? **A:** 8π **B:** 6π **C:** 2π **D:** -4π **E:** -6π

Answer: D

Model Response & Evaluation

Model `Ministral-3b-2512`: Invokes Gauss–Bonnet and writes $I = \int_M K \, d\sigma = 2\pi\chi(M)$. It then states that for orientable closed surfaces the Euler characteristic depends on the genus g, listing examples such as the sphere ($g = 0 \Rightarrow \chi = 2$) and the torus ($g = 1 \Rightarrow \chi = 0$), and also writes the general relation $\chi(M) = 2 - 2g$. It subsequently claims that the "double torus" ($g = 2$) still has $\chi(M) = 0$, and concludes that the only possible totals are $I = 4\pi$ (sphere) or $I = 0$ (all other orientable surfaces). Since neither 4π nor 0 appears in the options, it answers None.

Evaluation: Incorrect.✗
Error type: Reasoning Error. The failure occurs at the step where the model asserts "$g = 2$ still has $\chi(M) = 0$" and therefore "all non-spherical orientable surfaces have $\chi = 0$". This directly contradicts the model's own stated relation $\chi(M) = 2 - 2g$, so its elimination of option D is unjustified.

Model `Qwen3-vl-8b-instruct`: Applies Gauss–Bonnet $\int_M K \, d\sigma = 2\pi\chi(M)$. For a connected orientable closed surface of genus g, it uses $\chi(M) = 2 - 2g$, thus $I = 4\pi(1 - g)$. It notes $g = 2 \Rightarrow I = -4\pi$ and answers D.

Evaluation: Correct.✓
Analysis: Uses Gauss–Bonnet and $\chi(M) = 2 - 2g$ to derive $I = 4\pi(1 - g)$, then correctly matches $g = 2$ to option D.

Model `Claude-haiku-4.5`: Applies Gauss–Bonnet $\int_M K \, d\sigma = 2\pi\chi(M)$. For genus g, it uses $\chi(M) = 2-2g$, hence $I = 4\pi(1 - g)$. It checks the options via feasible Euler characteristics (even integers) and identifies $g = 2 \Rightarrow I = -4\pi$, answering D.

Evaluation: Correct.✓
Analysis: Correctly reduces the problem to topological invariants via Gauss–Bonnet, ruling out impossible $\chi(M)$ values and selecting $g = 2 \Rightarrow I = -4\pi$ (option D).

Model `Gemini-3-pro-preview`: Applies Gauss–Bonnet $\int_M K \, d\sigma = 2\pi\chi(M)$. For a connected orientable closed surface of genus g, it uses $\chi(M) = 2 - 2g$, thus $I = 2\pi(2 - 2g) = 4\pi(1 - g)$. Checking the options, it identifies $g = 2 \Rightarrow I = -4\pi$ and answers D.

Evaluation: Correct.✓
Analysis: Correctly applies Gauss–Bonnet with $\chi(M) = 2 - 2g$ and matches $g = 2$ to option D.

Figure 17: A differential topology problem together with representative model responses and error analysis. Correct answers and model answers are highlighted in red.

Problem in Algebraic Topology

Fill in the blank. Let S be the Klein bottle. Compute

$$H_1(S;\mathbb{Q}) \cong ____.$$

Answer: $\boxed{\mathbb{Q}}$

Model Response & Evaluation

Model `Ministral-3b-2512`: States that the Klein bottle is a closed surface of genus 2, and claims that

$$H_1(S;\mathbb{Q}) \cong \boxed{\mathbb{Q} \oplus \mathbb{Q}}.$$

It further asserts that the second homology group satisfies $H_2(S;\mathbb{Q}) \cong \mathbb{Q}$, treating the Klein bottle analogously to an orientable genus–2 surface.
Evaluation: Incorrect.✗
Error type: Knowledge Deficiency. The model lacks basic topological knowledge about the Klein bottle, incorrectly treating it as an orientable surface of genus 2. This leads to an erroneous identification of its homology groups, including both the rank of $H_1(S;\mathbb{Q})$ and the (nonexistent) top-dimensional rational homology.

Model `Qwen3-vl-8b-instruct`: Recalls the standard result $H_1(S;\mathbb{Z}) \cong \mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$. It then tensors with $\mathbb{Q}$, noting that the torsion summand vanishes, and concludes that

$$H_1(S;\mathbb{Q}) \cong \boxed{\mathbb{Q}}.$$

Evaluation: Correct.✓
Analysis: Correctly applies the universal coefficient theorem, observing that torsion in $H_1(S;\mathbb{Z})$ disappears over $\mathbb{Q}$, leaving a single free generator.

Model `Claude-haiku-4.5`: Derives $H_1(S;\mathbb{Z})$ by abelianizing the fundamental group $\pi_1(S) = \langle a, b \mid aba^{-1} = b^{-1} \rangle$, obtaining $\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$. Tensoring with $\mathbb{Q}$, it eliminates the torsion term and concludes

$$H_1(S;\mathbb{Q}) \cong \boxed{\mathbb{Q}}.$$

Evaluation: Correct.✓
Analysis: Correctly uses abelianization and the universal coefficient theorem to identify the rank–one rational homology group.

Model `Gemini-3-pro-preview`: Starts from the known integral homology $H_1(S;\mathbb{Z}) \cong \mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ and applies the universal coefficient theorem. It explicitly shows that $\mathbb{Z}/2\mathbb{Z} \otimes \mathbb{Q} = 0$, and concludes

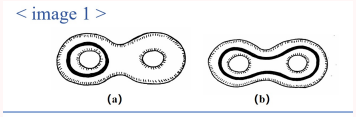
$$H_1(S;\mathbb{Q}) \cong \boxed{\mathbb{Q}}.$$

Evaluation: Correct.✓
Analysis: Correctly computes rational homology by eliminating torsion and identifying a single free generator.

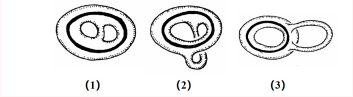
Figure 18: An algebraic topology problem on the Klein bottle together with representative model responses and error analysis. Correct answers and model answers are highlighted in red.

Problem in Geometric Topology

< image 1 >



< image 2 >



A circle is drawn on a pretzel with two holes (Image 1(a)). Show that it is possible to deform the pretzel so that the circle will be in the position represented in Image 1(b). Sort the figures in Image 2 to show this process.

Multiple Choice:
A: $(2)\to(1)\to(3)$ **B:** $(2)\to(3)\to(1)$ **C:** $(1)\to(2)\to(3)$ **D:** $(1)\to(3)\to(2)$ **E:** $(3)\to(2)\to(1)$

Answer: E

Model Response & Evaluation

Model `Ministral-3b-2512`: The sequence $(1)\to(2)\to(3)$ is interpreted as a continuous deformation of the pretzel. The transformation is described informally in terms of "pulling" and "stretching", without explicitly checking whether the intermediate states align with the start and end configurations shown in Image 1. The final choice is C.

Evaluation: Incorrect.✗
Error type: Image Understanding Error. The correspondence between Image 1(a)/(b) and the intermediate states in Image 2 is mismatched, leading to an invalid ordering of deformation steps.

Model `Qwen3-vl-8b-instruct`: The reasoning repeatedly revisits which panel represents the initial configuration and which represents the final one. Different assumptions are explored during the explanation, and the sequence is eventually fixed by treating (1) as the starting state and (3) as the target, without establishing a consistent global topological correspondence. This results in the selection C.

Evaluation: Incorrect.✗
Error type: Reasoning Error. The argument mentions local stretching and deformation, then loses a coherent global mapping across panels and terminates in an inconsistent sequence choice.

Model `Claude-haiku-4.5`: Image 1(a) is matched to state (3) and Image 1(b) to state (1). State (2) is treated as the necessary intermediate configuration in which one handle is pulled through the loop by a continuous deformation. This yields the ordering $(3)\to(2)\to(1)$ and the choice E.

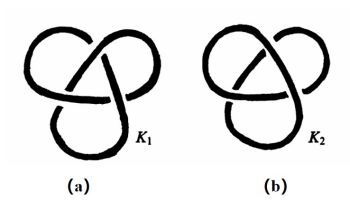
Evaluation: Correct.✓
Analysis: The global topological configuration is tracked consistently across all panels, and the intermediate deformation is used to justify the final ordering.

Model `Gemini-3-pro-preview`: State (3) is identified as the starting configuration and (1) as the target. State (2) is interpreted as the stage in which the right handle is pulled through the drawn circle during deformation. On this basis, the sequence $(3)\to(2)\to(1)$ is concluded, corresponding to E.

Evaluation: Correct.✓
Analysis: The ordering preserves a consistent mapping between the loop position and the deformed surface across states.

Figure 19: A topology deformation problem (pretzel with two holes) together with representative model responses and error analysis. Correct answers and model answers are highlighted in red.

Problem in Geometric Topology



As shown in the figure, which diagrams represent real knots (i.e., not unknots)?

Multiple Choice:
A: Only (a) **B:** Only (b) **C:** (a) and (b) **D:** None

Answer: A

Model Response & Evaluation

Model `Ministral-3b-2512`: Diagram (a) is identified as a trefoil knot, which is non-trivial. Diagram (b) appears visually untwistable into a simple loop. Selects A.

Evaluation: Correct.✓
Analysis: The conclusion is correct. The reasoning relies mainly on visual naming ("trefoil" vs. "untwistable") and lacks an explicit analysis of the over/under crossing structure or a concrete Reidemeister-move sequence for diagram (b).

Model `Qwen3-vl-8b-instruct`: Diagram (a) is a trefoil knot. Diagram (b) is interpreted as a figure-eight knot with four crossings. Selects C.

Evaluation: Incorrect.✗
Error type: Image Understanding Error. Diagram (b) has three crossings and is misidentified as a four-crossing figure-eight knot. The response misses the over/under structure that permits simplification to the unknot.

Model `Claude-haiku-4.5`: Diagram (a) matches a standard trefoil configuration. Diagram (b) can be deformed into a simple loop through continuous deformation. Selects A.

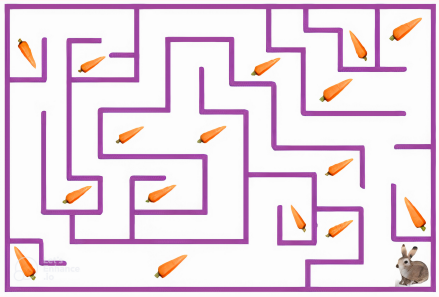
Evaluation: Correct.✓
Analysis: The classification is correct. The argument appeals primarily to known knot types and invariants. A complete justification would give the step-by-step Reidemeister simplification of diagram (b).

Model `Gemini-3-pro-preview`: Tracing the crossings in (b), one loop passes over the remaining strands twice consecutively. A Reidemeister type II move removes a pair of crossings. Selects A.

Evaluation: Correct.✓
Analysis: A valid Reidemeister type II cancellation is correctly identified. A residual kink remains after removing the crossing pair, which would require an additional Reidemeister type I move to fully reduce the diagram to the unknot. This final simplification step is left implicit.

Figure 20: A knot-identification problem with representative model responses and error analysis. Correct answers and model answers are highlighted in red.

Problem in Point-set Topology



A rabbit is placed in a maze containing several carrots. The rabbit can move freely along the corridors without crossing any walls. How many carrots can the rabbit reach and eat?

Multiple Choice:
A: 7 **B:** 8 **C:** 9 **D:** 15 **E:** 16

Answer: **B**

Model Response & Evaluation

Model `Ministral-3b-2512`: The maze is treated as largely open, and carrots are counted by visually grouping them into rows and regions (bottom, middle, top) without tracing the connectivity constraints imposed by the walls. Selects **E**.

Evaluation: Incorrect.✗
Error type: Image Understanding Error. The reasoning conflates *visible* carrots with *reachable* carrots and treats isolated chambers as accessible.

Model `Qwen3-vl-8b-instruct`: The solution attempts to enumerate carrots sequentially by imagining step-by-step movement through the maze. The traversal repeatedly restarts and revises earlier counts, leading to inconsistent intermediate totals. Selects **None**.

Evaluation: Incorrect.✗
Error type: Reasoning Error. The model recognizes that wall constraints matter, then loses a stable global notion of reachability. Repeated recounting without a fixed connectivity map produces inconsistent totals.

Model `Claude-haiku-4.5`: All carrots are first counted by partitioning the maze into upper, middle, and lower regions. The corridors are then assumed to form a single connected component. Selects **D**.

Evaluation: Incorrect.✗
Error type: Image Understanding Error. The reasoning distinguishes spatial regions correctly and assumes global connectivity incorrectly. Several carrots lie inside fully enclosed subregions that cannot be reached from the rabbit's starting position.

Model `Gemini-3-pro-preview`: To solve this puzzle, we need to trace the path the rabbit can take from its starting position in the bottom right corner and count every carrot it can reach without crossing any purple lines. Starting at the rabbit in the bottom right, we move up and encounter a carrot in the vertical corridor (Count: 1). From there, moving left and then down leads to a dead end containing another carrot (Count: 2)... Selects **C**.

Evaluation: Incorrect.✗
Error type: Image Understanding Error. The claimed dead end and additional carrot are absent from the indicated route in the figure.

Figure 21: A maze-reachability problem with representative model responses and error analysis. Correct answers and model answers are highlighted in red.

Problem and Outcome



Task. The five drawings show closed flexible-band configurations A–E. The model must identify the configuration whose knot type prevents transformation into the others through continuous deformation.

Expert-adjudicated answer: E **OpenAI o3 answer after repeated inspection:** A

Five-Turn Inspection Trajectory and Failure Analysis

Protocol. The interaction provides up to five image-inspection turns. Each turn records the inspected region, visual evidence, remaining uncertainty, and the next planned inspection. The final turn requests an answer, brief justification, and confidence.

Observed trajectory.

- **Turn 3:** o3 classifies C–E as unknots and initially treats A and B as trefoil-like mirror configurations. It plans to resolve their chirality.
- **Turn 4:** after recounting visible crossings, it reclassifies B as an unknot, groups B–E together, and selects A as the unique knot.
- **Turn 5:** a final side-by-side inspection repeats the same crossing-count argument and again selects A, with reported confidence 0.73.

Failure mechanism. The response substitutes apparent crossing count for global knot type. The number of crossings visible in a particular drawing is insufficient: local strands must be followed consistently, and candidate simplifications must preserve the complete over/under structure through valid deformations.

The trajectory also changes B from a trefoil-like configuration in Turn 3 to an unknot in Turn 4 without a corresponding deformation argument. Subsequent inspection revisits the same local cue and preserves the incorrect answer.

Case-level conclusion. The adjudicated answer E requires a stable representation of the entire closed band and its continuous untangling path. In this case, additional image inspections increased observation time while retaining an incorrect global topological representation.

Figure 22: **OpenAI o3 think-with-image failure case.** The five-turn protocol elicits repeated visual inspection, a changing classification of B, and a final answer driven by apparent crossing counts. The expert-adjudicated answer is E; o3 selects A. This diagnostic trajectory is not included in the aggregate benchmark comparisons.