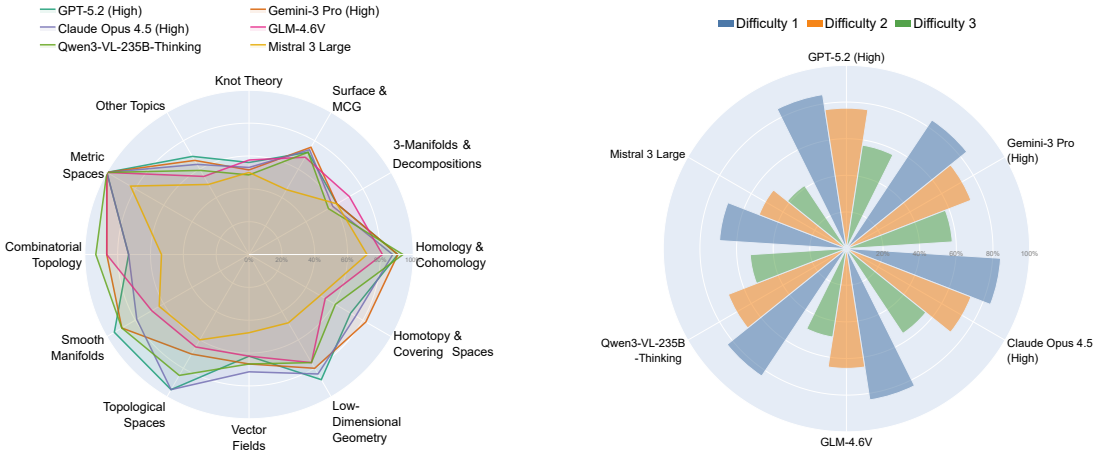
TopoEval: A Comprehensive Benchmark for Topological Reasoning in Foundation Models

Anonymous ACL submission



(a) Performance across 12 fine-grained topology subfields.

(b) Performance across 3 difficulty levels.

Figure 1: **Results of 6 representative Foundation Models.** The parentheses in the model names indicate the level of inference-time reasoning effort. (a) Model performance across fine-grained topology subfields. (b) Performance across 3 difficulty levels: Level 1 (easiest) to Level 3 (most difficult).

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 often treat topology superficially, lacking systematic coverage and depth. To bridge this gap, we introduce TopoEval, a meticulously curated topological reasoning benchmark comprising 400 high-quality problems adapted from real-world math competitions and professional textbooks. TopoEval features a rigorous hierarchical taxonomy, spanning 4 major topological branches subdivided into 12 fine-grained subfields, and is graded across 3 difficulty levels. Incorporating tasks that demand complex visual reasoning, our dataset presents a comprehensive challenge to foundation models. Through extensive experimentation, we systematically investigate the impact of model scaling, reasoning depth, and prompt engineering strategies on performance. Furthermore, detailed error analysis unveils the deficiencies of current models in topological reasoning, providing critical directions for developing systems with genuine visual understanding and reasoning capabilities.

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 significantly in geometric appearance, they are considered equivalent in a topological sense because both contain exactly one "hole" (as illustrated in 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 increasingly become the central pillars of modern artificial intelligence systems, a fundamental question arises: **Do current state-of-the-art foundation models exhibit topological understanding comparable to that of humans?**

The dataset is available at https://topoeval-site.pages.dev/.

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.

Recent progress in foundation models has led to the emergence of increasingly capable systems, including the open-source Qwen3 (Alibaba Cloud AI Team, 2025), the open-weight LLaMA-4 (Meta AI, 2025), proprietary GPT-5.2 (OpenAI, 2025b) and Gemini-3 (Google DeepMind, 2025). These modern models simultaneously integrate advanced reasoning mechanisms and 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., 2023a). 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 problems in topology can be viewed as visual reasoning problems.

General-purpose visual reasoning datasets such as MathVista (Lu et al., 2024), MMMU (Yue et al., 2024), and MathVision (Wang et al., 2024) **are not designed as dedicated benchmarks for topological visual reasoning**, let alone curate questions spanning diverse subfields of topology. Furthermore, the few topology-related problems included in these datasets are often superficial, exhibiting limited difficulty and depth. This scarcity of specialized, high-quality data prevents a precise diagnosis of model failures in understanding complex topological invariants.

To address this gap, we introduce **TopoEval**, **a benchmark specifically designed to evaluate topological reasoning in foundation models**. TopoEval comprises 400 carefully curated topology problems, which systematically cover multiple subfields such as point-set topology, geometric topology, algebraic topology, and differential topology, and is further divided into 12 fine-grained subfields that capture most core topological concepts and methods. It spans academic levels ranging from K12 to professional training. A substantial portion of these problems requires reasoning over visual representations, with fine-grained annotations provided for dimensions such as image dependency, visual complexity, and problem difficulty, thereby enabling a multi-dimensional analysis of models' topological reasoning capabilities. We note that TopoEval contains 400 problems, which is smaller than some large-scale multimodal benchmarks (Liu et al., 2023c; Li et al., 2023b). This difference arises because many TopoEval items involve specialized topological concepts and require expert-level adaptation from competition problems and professional textbooks, making curation and verification substantially more time-consuming than generic data collection. Indeed, in expert-curated evaluations of advanced mathematics, benchmark sizes are often similarly modest (Zheng et al., 2021; Azerbayev et al., 2023; Poiroux et al., 2025).

Based on TopoEval, we conducted a large-scale empirical study across diverse model settings, systematically evaluating a range of open-source and closed-source models. We further investigated the effects of reasoning depth and prompt engineering strategies on model performance. Overall, even under the optimal model configuration, the highest accuracy achieved on the full dataset is only 73.00%. The average accuracy of open-source and closed-source models reaches 55.67% and 64.99%, respectively, both substantially lower than the 86.60% accuracy achieved by human experts. **These results indicate that topological reasoning remains a significant and unresolved challenge for current foundation models**. Notably, models exhibit a pronounced deficiency on K12 level topology problems. Although these problems do not rely on complex specialized knowledge, they depend heavily on intuitive understanding and imagination of concepts such as knot and continuous deformation. Even the strongest models perform substantially worse than human experts on this subset, indicating a clear gap between current models and humans

at the most fundamental and intuitive level of topological understanding. In addition, model performance is markedly weaker in scenarios with strong reliance on visual information. As visual complexity or image dependency increases, model accuracy shows a clear downward trend. Importantly, this trend persists even when larger model scales or higher levels of reasoning effort are employed.

Our main contributions are threefold:

- **Benchmark Construction:** We introduce TopoEval, which, to the best of our knowledge, is the first benchmark specifically designed to evaluate topological reasoning in foundation models. It systematically covers multiple topological subfields, varying difficulty levels, and academic tiers ranging from K12 to professional.
- **Systematic Evaluation:** We conduct a comprehensive empirical evaluation of a wide range of mainstream foundation models and their inference settings, revealing that topological reasoning remains a persistent challenge even for state-of-the-art systems.
- **Analysis and Insights:** Through fine-grained analysis and case studies, we uncover systematic weaknesses in current foundation models regarding intuitive topological reasoning and visual structural understanding, providing new empirical evidence for their limitations in abstract structural cognition.

2 Related works

2.1 From Training-Time Scaling to Inference-Time Reasoning

Long-standing improvements in foundation models have largely followed the *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). However, as model scales continue to grow, the cost of acquiring sufficiently large amounts of high-quality training data rises substantially, while marginal returns increasingly diminish (Villalobos et al., 2022; Chen et al., 2025). As a result, relying solely on training-time scaling is approaching practical bottlenecks.

Against this backdrop, research attention has been shifting toward *inference-time scaling*, which aims to improve model performance by allocating 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,c). Building on a base model, such systems typically combine supervised fine-tuning with reinforcement-learning-based post-training, leveraging high-quality data to elicit stronger multi-step reasoning behaviors and enhanced search over reasoning paths (Ouyang et al., 2022; Bai et al., 2022).

Despite these advances, the sustainability of inference-time scaling faces serious challenges. Inspecting models' thought traces often reveals redundant and unproductive reasoning, where excessive deliberation increases computational cost and may even harm accuracy on intuition-heavy problems (Snell et al., 2024; Anthropic, 2025c).

Many problems in TopoEval benefit from geometric intuition for efficient reasoning-path search, making the benchmark well suited to evaluate whether foundation models leverage such intuition in topological reasoning, rather than relying on indiscriminate increases in reasoning depth.

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., 2023b). 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 effectively exploit visual evidence, instead relying on language priors acquired during pre-training (Li et al., 2023c; Liu et al., 2024).

In reasoning tasks governed by strong geometric properties, such as topology (Munkres, 2000), issues of visual grounding and hallucination are further amplified. Unlike general visual reasoning tasks that primarily emphasize object recognition or coarse-grained relationships, topological problems (e.g., analyzing mazes, knots, or continuous deformations (Adams, 2004)) involve core struc-

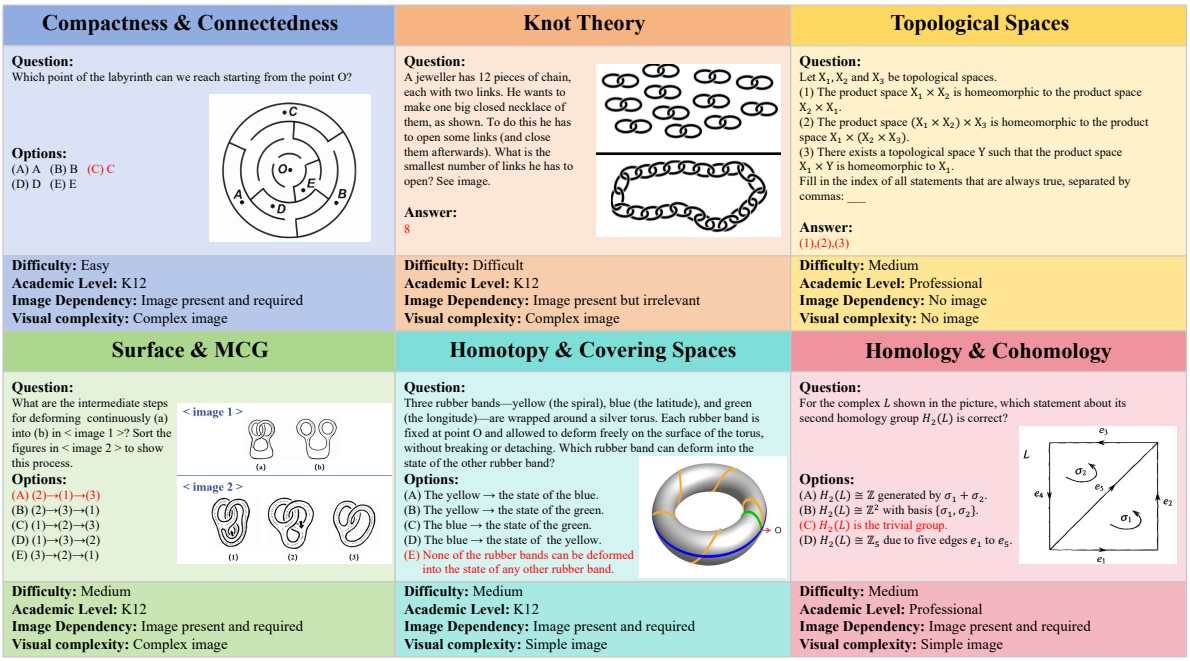


Figure 3: **Example problem cards from TopoEval.** Correct answers are highlighted in red. To facilitate understanding for a general audience, we intentionally select a larger proportion of K12 level problems here, which does not reflect the true distribution of the full dataset. Detailed dataset statistics are provided in Section 3.1.

tures composed of irregular curves and complex connectivity. Solving such problems requires precise discrimination of geometric forms: even subtle local visual discrepancies may correspond to fundamentally different entanglement patterns or connectivity structures, leading to divergent topological conclusions.

As a benchmark centered on topological reasoning, TopoEval places stringent demands on a model's ability to capture fine-grained visual details during inference. Furthermore, TopoEval introduces fine-grained evaluation dimensions, including image dependency and Visual Complexity. These dimensions enable a systematic analysis of performance trends with respect to visual difficulty, allowing researchers to disentangle whether observed performance gains arise from improved textual reasoning or from genuine advances in visual perception.

Table 1: **Dataset statistics of TopoEval.** Question length excludes answer options for multiple-choice questions.

Statistic	Number
Total questions	400
Question format	
– Multiple-choice questions	324 (81.0%)
– Free-form questions	76 (19.0%)
Academic Level	
– Professional	315 (78.8%)
– K12	85 (21.2%)
Image dependency	
– No image	83 (20.8%)
– Image present but irrelevant	57 (14.2%)
– Image present and required	260 (65.0%)
Visual complexity	
– No image	83 (20.8%)
– Simple image	158 (39.5%)
– Complex image	159 (39.8%)
Question length (words)	
– Maximum	188
– Minimum	5
– Average	34.27

3 Datasets

3.1 TopoEval Overview

TopoEval is a curated benchmark for evaluating topological reasoning in foundation models. It contains 400 problems adapted from textbooks, lecture notes, and competition sources. Table 1 summarizes the dataset statistics, including the distribution of question formats, difficulty levels, visual attributes, and question length.

TopoEval provides a fine-grained characterization of visual attributes. Specifically, we categorize image dependency into three levels: no image, image present but irrelevant to solving the problem, and image present and required for reaching the correct solution. In addition, we explicitly annotate

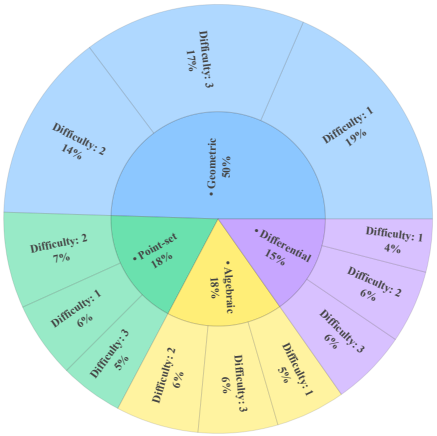


Figure 4: **Branch–difficulty distribution of TopoEval.** Inner ring: four major topological branches; outer ring: three difficulty levels.

visual complexity, distinguishing among problems with no image, simple images, and complex images. These fine-grained annotations enable systematic analysis of whether models can effectively integrate textual and visual information during reasoning, as well as whether they can accurately perceive and utilize fine-grained details in images to support correct topological reasoning.

Each problem is categorized into a hierarchical taxonomy. At the top level, TopoEval spans four major branches—*Point-set*, *Algebraic*, *Geometric*, and *Differential*. Figure 4 visualizes the branch–difficulty distribution, showing broad coverage across branches and a balanced spread over difficulty levels. In addition, the dataset further subdivides the four major branches into twelve fine-grained topological subfields. For details, see 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* (Ye, 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 leverage foudation 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, and adapt selected items into evaluable multiple-choice or fill-in formats following the curation protocol in Section 3.3.

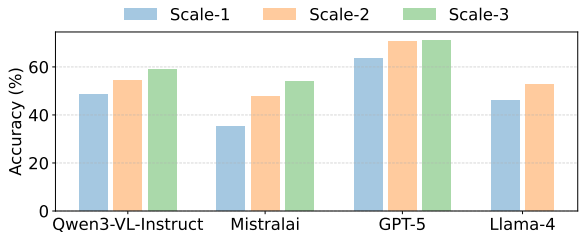
3.3 Data Adaptation and Curation

Many topology problems in their original form are presented as proof-based questions. While such problems are fundamental to mathematical research and education, they pose substantial challenges for automated evaluation. One possible approach is to formalize these problems and verify solutions using interactive theorem provers such as Lean or Isabelle (de Moura et al., 2015; Nipkow et al., 2002). However, automatic formalization and automated proof techniques remain at an early stage, and have been shown to struggle with informal mathematics, visual inputs, and intuition-heavy reasoning, making them unsuitable for reliably handling topology problems of the kind considered in this benchmark (Avigad et al., 2014; Chen et al., 2021). We therefore do not pursue this direction in this work.

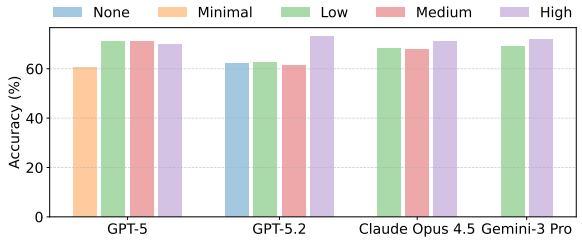
To ensure evaluability and reproducibility, we systematically adapt proof-based problems into multiple-choice or fill-in-the-blank formats. Moreover, many studies have shown that, under such question type, automated judging with foundation models achieves near-perfect accuracy in answer extraction and verification (Lu et al., 2024; Wang et al., 2024). In addition, reformulating problems into these formats helps mitigate evaluation set leakage in foundation model assessment, as the adapted questions are less likely to be directly memorized or matched from existing solution corpora (Chen et al., 2021; Zhou et al., 2023).

During the adaptation stage, we employ an LLM-assisted annotation and adaptation system. The system produces structured annotations covering candidate answers, fine-grained topological subfields, image dependency, visual complexity, academic level, and difficulty estimates. All outputs are subsequently manually reviewed and validated by at least one graduate student with formal training in topology, who confirms the correctness of the final answers and resolves any ambiguities introduced

during automated adaptation. A detailed description of the design of this annotation–adaptation system is provided in Appendix B. An online demonstration of the annotation system is available at https://annotation1103.pages.dev/.



(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.** Please refer to Table 4 for the exact model scale mappings.

4 Experiments

In this section, we present a unified evaluation of a diverse set of representative foundation models on TopoEval and analyze how model families and inference configurations affect topological reasoning performance. We first report overall accuracy and fine-grained results across the 12 topology subfields, and then examine scaling trends along two axes: parameter-side scaling and inference-time scaling (i.e., varying computation budgets at test time). Next, we analyze performance as a function of visual difficulty, focusing on image dependency and visual complexity. We further compare results across academic tiers, contrasting K12 problems with professional-level ones. Finally, we study prompt-based inference strategies, including Chain-of-Thought prompting and multi-sample aggregation (Wei et al., 2022; Snell et al., 2024); however, under our setting, both strategies yield unstable gains (Tables 5 and 6).

4.1 Experimental Setup

Following prior work on large-scale foundation model evaluation, we select representative models based on 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 Qwen3-VL family (Alibaba Cloud AI Team, 2025), LLaMA-4 variants (Meta AI, 2025), GLM-4.6V (Zhipu AI, 2025), ERNIE-4.5-VL (Baidu Research, 2025), and the Mistral/Ministral series (Mistral AI, 2025). Our proprietary set consists of state-of-the-art systems from major providers, including GPT-5 and GPT-5.2 (OpenAI, 2025b), the Gemini-3 series (Google DeepMind, 2025), the Claude-4.5 series (Anthropic, 2025b), and Grok-4 and Grok-4.1 (xAI, 2025a).

Unless otherwise specified, all models are evaluated in a zero-shot setting. Model outputs are judged by GPT-5-nano (OpenAI, 2025b), which serves as a unified judge model for answer extraction, normalization, and correctness verification. We randomly inspected 100 automatically evaluated questions, and only one item was incorrectly judged due to a failure in correctly extracting the final answer. All experiments are conducted through the OpenRouter API (OpenRouter, 2024), which provides a unified interface for accessing both open-weight and proprietary models. Unless otherwise specified, the temperature parameter is set to 0 and the Top-p is set to 0.75.

4.2 Experimental Results

Overall results. Figure 1 provides a high-level overview of model performance across fine-grained topology subfields and difficulty levels, while Table 2 reports detailed accuracy results across all 12 subfields. Overall, proprietary models substantially outperform open-weight models, yet a clear gap to human experts remains across both aggregate and fine-grained evaluations. In terms of overall accuracy, the strongest proprietary model is GPT-5.2 (high) at 73.0%, while the best open-weight model is Qwen3-VL-235B-Thinking at 70.0%. By comparison, human experts achieve 86.6% accuracy, leaving a 13.6-point gap to the best model. Across fine-grained subfields, we observe consistent performance disparities. Overall, models perform better on subfields with lower image dependency and lower visual complexity. Specifically, models perform strongest on *Metric Spaces* and *Homology*

Table 2: **Performance comparison across 12 fine-grained topology subfields.** We report accuracy (%) for representative models. Red indicates the best performance, and Blue indicates the second best (excluding Human Experts). The parentheses in the model names indicate the level of inference-time reasoning effort. See Table 7 for the definitions of the 12 fine-grained subfields. Results for the remaining models are reported in Table 8.

Model	Overall	Knot	Surf	3M	Hom	Hty	LDG	VFld	TopSp	SmMfd	Comb	Metric	Other
Random Chance	17.9	13.8	26.3	26.5	15.4	14.8	16.0	26.7	22.2	10.0	16.7	28.6	22.6
Proprietary Models													
GPT-5 (minimal)	60.8	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)	70.0	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-5.2 (none)	62.0	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)	73.0	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	70.8	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)	71.8	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)	66.0	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)	71.0	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	65.5	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	66.5	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	59.0	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	70.0	50.5	68.4	70.6	93.8	82.1	84.0	76.2	93.8	86.8	86.7	100.0	67.6
GLM-4.6V	66.0	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	53.0	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	54.5	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	52.5	31.3	50.9	41.2	84.4	50.0	72.0	47.6	83.8	68.4	43.3	94.1	45.1
Human Experts	86.6	82.5	88.0	79.5	98.0	92.0	95.0	85.0	99.0	95.0	90.0	100.0	81.5

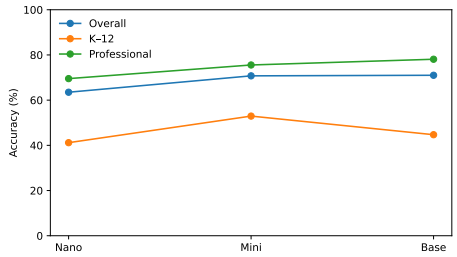
& Cohomology. In contrast, the most challenging subfields are *Knot Theory* and *3-Manifolds & Decompositions*. In Appendix E, we conduct a case study on a carefully selected, representative subset of problems from the dataset, where we categorize common error types and further examine whether the reasoning processes of models are correct even when their final answers are correct.

Parameter-side vs. inference-time scaling. Figures 5a and 5b compare parameter-side scaling and inference-time scaling. The results show that both increasing model size and allocating more computation at inference time have the potential to improve model performance. However, for some model families (e.g., GPT-5), both scaling strategies appear to gradually encounter bottlenecks. A closer inspection of Figure 5b further reveals that, for topological reasoning tasks, the performance gains from continuously increasing inference-time reasoning effort are not stable. This is likely because certain topological reasoning problems rely more heavily on intuitive judgment and a thorough understanding of visual information, rather than simply extending the length of the reasoning chain.

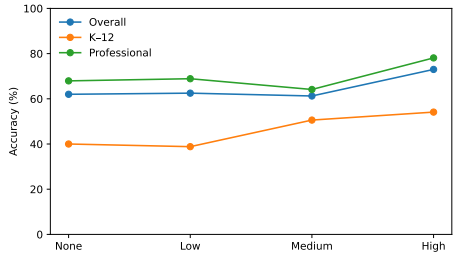
Image dependency and complexity. Table 3 shows that model performance is highly sensitive to both image dependency and visual complexity. Here, Non-Ima denotes questions without images; Dep-1 denotes questions with an image that is present but not required; Dep-2 denotes image-required questions; and Com-1/Com-2 denote questions with simple/complex images. Across all model families, accuracy on Non-Ima and Dep-1 is generally higher than the overall accuracy, with relative gains ranging from -1.6% to +31.9% (Non-Ima) and +6.6% to +43.7% (Dep-1). In contrast, performance degrades systematically on Dep-2, with relative drops of about 5.7%–17.2%, indicating persistent difficulty when visual evidence is essential. A similar pattern holds for visual complexity: Com-1 has a small effect for most models, whereas Com-2 causes substantial degradation across the board, with relative drops ranging from 5.8% to 24.5% and often exceeding 15%. Overall, these results suggest that failures on TopoEval are primarily driven by deficiencies in vision-critical and structurally complex cases, rather than insufficient textual reasoning alone.

Table 3: **Performance under image dependency and visual complexity.** For each model, the first row reports raw accuracy (%). The second row reports the relative change w.r.t. Overall (%). Green/red indicates improvement/degradation.

Model	Overall	Non-Ima	Dep-1	Dep-2	Com-1	Com-2
GPT-5 (minimal)	60.8	61.4	78.9	56.5	65.8	55.3
		+1.1%	+30.0%	-6.9%	+8.4%	-8.9%
GPT-5 (high)	70.0	84.3	86.0	61.9	70.3	62.3
		+20.5%	+22.8%	-11.5%	+0.4%	-11.1%
GPT-5.2 (none)	62.0	63.9	75.4	58.5	65.8	57.2
		+3.0%	+21.7%	-5.7%	+6.2%	-7.7%
GPT-5.2 (high)	73.0	89.2	82.5	65.8	73.4	64.2
		+22.1%	+13.0%	-9.9%	+0.6%	-12.1%
Gemini-3 Flash	70.8	85.5	75.4	65.0	73.4	60.4
		+20.9%	+6.6%	-8.1%	+3.8%	-14.7%
Gemini-3 Pro (high)	71.8	86.7	86.0	63.8	74.1	61.6
		+20.9%	+19.8%	-11.0%	+3.2%	-14.1%
Claude Opus 4.5 (low)	68.2	84.3	84.2	59.6	73.4	54.7
		+23.6%	+23.4%	-12.7%	+7.6%	-19.8%
Claude Opus 4.5 (high)	71.0	88.0	84.2	62.7	75.9	57.2
		+23.9%	+18.6%	-11.7%	+7.0%	-19.4%
Grok-4 Fast	64.0	84.3	82.5	53.5	63.9	53.5
		+31.8%	+28.8%	-16.5%	-0.1%	-16.5%
Grok-4	66.5	85.5	89.5	55.4	67.1	56.0
		+28.6%	+34.5%	-16.7%	+0.9%	-15.8%
Qwen3-VL-235B-Instruct	58.8	57.8	78.9	54.6	62.7	55.3
		-1.6%	+34.4%	-7.0%	+6.7%	-5.8%
Qwen3-VL-235B-Thinking	68.5	90.4	75.4	60.0	69.0	56.6
		+31.9%	+10.1%	-12.4%	+0.7%	-17.4%
GLM-4.6V	66.0	75.9	82.5	59.2	70.9	56.0
		+15.0%	+24.9%	-10.3%	+7.4%	-15.2%
ERNIE-4.5-VL	52.8	60.2	71.9	46.2	59.5	42.1
		+14.2%	+36.4%	-12.5%	+12.8%	-20.1%
Mistral-Large	54.0	62.7	71.9	47.3	58.2	45.3
		+16.0%	+33.2%	-12.4%	+7.8%	-16.1%
LLaMA-4 Maverick	52.5	65.1	75.4	43.5	58.9	39.6
		+23.9%	+43.7%	-17.2%	+12.1%	-24.5%



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



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

Figure 6: **Academic-tier performance: K12 vs. Professional (%) in different settings.**

K12 vs. Professional. Figure 6 compares model performance on K12 and professional-level problems and reveals a consistent tier gap under both inference-time scaling and parameter-side scaling. Taking GPT-5.2 as an example, as inference-time reasoning effort increases, overall accuracy rises from about 60% to about 70%, while accuracy on professional-level problems improves from roughly 68% to nearly 80%. In contrast, accuracy on K12 problems remains substantially lower, increasing only from about 40% to about 55%, and exhibits noticeable fluctuations at intermediate effort levels, indicating unstable performance gains. A similar pattern is observed under parameter-side scaling. For GPT-5, moving from smaller to medium-sized models improves both overall and K12 performance, but further increasing model size leads to a decline in K12 accuracy (from about 50% to about 45%), while professional-level accuracy continues to rise and approaches 80%. Overall, performance on professional-level problems scales more reliably with both model size and inference-time computation, whereas K12 problems remain a persistent bottleneck even with larger models or higher reasoning budgets. A key reason for this disparity is that K12 problems rely more heavily on intuitive and perceptual visual reasoning rather than formalized domain knowledge. As a result, simply increasing a model's knowledge capacity is often insufficient to improve accuracy on such problems. Moreover, K12 problems in TopoEval exhibit substantially higher image dependency and visual complexity than professional-level problems. This heightened reliance on visual information makes it difficult for models to compensate through increased parameter scale or inference-time reasoning effort alone.

5 Conclusion

We introduce TopoEval, a curated benchmark for systematically evaluating topological reasoning in foundation models. TopoEval covers major topological branches and fine-grained subfields, with annotations that characterize image dependency, visual complexity, and problem difficulty. Our results show that topological reasoning remains challenging for current foundation models. Errors are concentrated in cases requiring strong visual grounding and intuitive structural understanding, particularly on K12-level problems. These findings indicate that current models still struggle to form stable, structure-preserving interpretations from visual inputs.

We hope that TopoEval will serve as a useful diagnostic benchmark for future research, facilitating more targeted investigations into visual grounding, intuitive reasoning and topological understanding.

Limitations

Exclusion of Formal Verification. Although automated formalization and theorem proving are active research areas in AI for mathematics, we did not use formal languages like Lean or Isabelle to construct the dataset. Proof assistants such as Lean, Coq, and Isabelle are widely used to formalize mathematics rigorously, but many problems in TopoEval rely on visual reasoning and intuition that current formalization techniques struggle to capture reliably (Naskręcki, 2024). Moreover, automatic translation of informal mathematical text into a formal language remains a challenging task, especially for higher-level concepts typical of topology (Patel et al., 2023).

Exclusion of Proof Generation Tasks. To ensure scalable and reproducible evaluation, TopoEval primarily uses multiple-choice and short-answer formats. Consequently, we do not assess a model's ability to generate rigorous mathematical proofs, even though constructing and verifying such proofs is a core competency in professional topology. Given that automated formalization and proof verification techniques are still in their early stages for complex mathematics, we opted for this simplified approach.

Absence of Topological Transformation-Based Data Synthesis. We did not use topological transformations, such as homeomorphisms, to batch-synthesize dataset instances, although such techniques could potentially generate synthetic problems and help alleviate dataset leakage (Ramos et al., 2025). This type of synthesis involves complex computational geometry and deep topological reasoning, which is beyond the scope of the current work (Hristov et al., 2025).

Ethical Statement

All questions in TopoEval are sourced from public educational and academic resources. This research strictly adheres to the objectives of non-profit educational research and foundation model evaluation, ensuring that data usage complies with the "Fair Use" principles of the academic community. We solemnly declare our opposition to the use of this dataset or associated models for commercial profit, academic dishonesty, or any other scenarios that violate ethical standards.

During the data dissemination process, we uphold rigorous copyright protection principles:

- **Image Resource Management:** We only publicly release images with explicit open-source licenses (e.g., Public Domain or Creative Commons). Images with ambiguous or restricted copyright status have been excluded from the public dataset; for these cases, only the question text, standard answers, and structured metadata are retained.
- **Content Reconstruction:** To prevent the direct reproduction of original materials, all questions have undergone substantial rewriting and paraphrasing. Furthermore, visual information has been converted into equivalent abstract representations. Through these de-identification and reconstruction techniques, we ensure that the publicly released content remains independent and irreversible, fully respecting the intellectual property rights of the original authors.

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*. ArXiv:2204.14198.

Alibaba Cloud AI Team. 2025. Qwen3 technical report. Technical Report.

Anthropic. 2025a. Claude opus 4.5. Product documentation.

Anthropic. 2025b. Claude opus 4.5 model card. Model Card.

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

Jeremy Avigad, John Harrison, and 1 others. 2014. The mechanization of mathematics. *Notices of the American Mathematical Society*, 61(6).

Zhangir Azerbayev, Bartosz Piotrowski, Jeremy Avigad, and 1 others. 2023. Proofnet: Autoformalizing and formally proving undergraduate-level mathematics. *arXiv preprint arXiv:2302.12433*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Zachary Goldie, and 1 others. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Baidu Research. 2025. Ernie-4.5-vl: A unified vision-language foundation model. Technical report, Baidu. Technical report.

Caribou Mathematics Contest. 2025. Caribou contests. https://cariboutests.com/. Accessed: 2026-01-05.

Jeremy Chen and 1 others. 2021. Evaluating mathematical formalization with natural language processing. In *Proceedings of the Conference on Automated Deduction (CADE)*.

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.

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, 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. In *Proceedings of the 25th International Conference on Automated Deduction (CADE)*.

Google DeepMind. 2025. Gemini 3 technical report. Technical Report.

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 arXiv:2203.15556*.

Petar Hristov, Ingrid Hotz, and Talha Bin Masood. 2025. Robust geometric predicates for bivariate computational topology. *arXiv preprint*. Highlights challenges in implementing geometric algorithms with exact arithmetic and robustness issues.

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 arXiv:2001.08361*.

John M. Lee. 2012. Simply connected spaces. https://www.math.washington.edu/~lee/Courses/441-2012/simplyconn.pdf. 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.

Wei Li and 1 others. 2023b. Seed-bench: Benchmarking multimodal large language models. *arXiv preprint arXiv:2307.16125*.

Yifan Li and 1 others. 2023c. POPE: Polling-based object probing evaluation for object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

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. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Haotian Liu and 1 others. 2023b. Visual instruction tuning. *Advances in Neural Information Processing Systems*.

Yuan Liu and 1 others. 2023c. Mmbench: Is your multimodal model an all-round player? *arXiv preprint arXiv:2307.06281*.

Pan Lu, Swaroop Mishra, Tony Xia, and 1 others. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Math Kangaroo Association. 2025. Math kangaroo. https://mathkangaroo.org/mks/. Accessed: 2026-01-05.

Meta AI. 2025. Llama 4: Open foundation and fine-tuned chat models. Model Release.

Mistral AI. 2025. Mistral and ministral: Efficient and scalable open foundation models. Technical report, Mistral AI. Technical report.

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

James R. Munkres. 2000. *Topology*. Prentice Hall.

B. N. Naskręcki. 2024. Mathematical discovery in the age of artificial intelligence. *Nature Physics*. Discusses AI-assisted formalization with proof assistants such as Lean, Coq, and Isabelle.

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

OpenAI. 2025a. Gpt-5.1. https://openai.com/research. Accessed: 2026-01-05.

OpenAI. 2025b. Gpt-5.2 system card. System Card.

OpenRouter. 2024. Openrouter: A unified interface for large language models. https://openrouter.ai. Accessed: 2025-01.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, and 1 others. 2022. Training language models with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Nilay Patel, Rahul Saha, and Jeffrey Flanigan. 2023. A new approach towards autoformalization. *arXiv preprint*, arXiv:2310.07957. Explores challenges in autoformalization of mathematical proofs.

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 17958–17980. Association for Computational Linguistics.

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. Data leakage occurs when evaluation images overlap with training data, leading to inflated performance metrics.

Charlie Snell, Katherine Lee, Ankur Kumar, and 1 others. 2024. Scaling test-time compute for large language models. *arXiv preprint arXiv:2402.XXXX*.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*.

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 arXiv:2211.04325*.

Zhen Wang and 1 others. 2024. Mathvision: Evaluating mathematical reasoning of multimodal large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*.

xAI. 2025a. Grok-4 technical overview. Technical Overview.

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

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 arXiv:2405.00451*.

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

Youcheng Ye. 1997. *Lecture Notes on Basic Topology*. Peking University 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.

Xiang Yue and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding benchmark. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kunhao Zheng, Jesse Michael Han, S. Zhou, and et al. 2021. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. *arXiv preprint arXiv:2109.00110*.

Zhipu AI. 2025. Glm-4.6v: Advancing multimodal foundation models with enhanced visual reasoning. Technical report, Zhipu AI. Technical report.

Wenhao Zhou and 1 others. 2023. Large-scale evaluation of neural language models for mathematical reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Nomenclature and Mappings

This section provides auxiliary nomenclature and mappings used throughout the paper to ensure clarity, consistency, and reproducibility. Specifically, Table 4 defines the correspondence between discrete parameter-side scaling indices and concrete model variants within each model family, which is used to unify model size comparisons across different architectures. Table 7 summarizes the abbreviations, full names, and scopes of the 12 fine-grained topology subfields in TopoEval, enabling concise presentation and interpretation of fine-grained results.

Table 4: **Parameter-side scaling mappings.** Within each model family, `scale` is a discrete index from smaller to larger parameter sizes.

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

B LLM-assisted Annotation and Adaptation System

This appendix describes the LLM-assisted annotation and adaptation system used to curate and standardize problems in TopoEval. The system is designed to support efficient human–AI collaboration during dataset construction, while ensuring structural consistency, annotation reliability, and reproducibility. An online demonstration of the annotation system is available at https://annotation1103.pages.dev/. Figure 7 illustrates the overall interface and interaction flow of the LLM-assisted annotation system, highlighting the structured workflow and human-in-the-loop design.

B.1 Design principles

The system is guided by three high-level design principles. First, it follows a *local-first* paradigm: all intermediate drafts, annotations, and assets are created and stored locally by default, allowing annotators to retain full control over data and reducing unintended information leakage. Second, the system enforces *schema-constrained generation*, requiring LLM outputs to conform to a predefined structured record format rather than free-form text. Third, it emphasizes *iterative reliability*, coupling generation with automatic review and normalization steps to mitigate common failure modes such as inconsistent answers, malformed notation, or rendering errors.

B.2 Structured problem representation

Each problem instance is represented as a structured record

$$r = \big(q,\ t,\ \mathcal{O},\ a,\ m,\ \pi\big),$$

where q denotes the question text, $t \in \{\mathrm{MCQ}, \mathrm{FIB}, \textsc{Proof}\}$ is the target question type, $\mathcal{O} = \{o_i\}_{i=1}^{n}$ is the option set for multiple-choice questions, a is the answer string, and m is metadata such as topological subfield, source type, academic level, and difficulty. The field π optionally references an associated image. In addition, we annotate whether the semantics of the problem depend on the image, enabling later analysis of visual dependency.

This structured representation allows annotation, validation, and export processes to be handled uniformly across diverse problem types.

B.3 LLM interaction and role specialization

The system adopts a role-specialized interaction pattern, in which different LLM agents are assigned distinct responsibilities within the annotation workflow. These roles include text transcription from images, LaTeX normalization, problem generation or completion, structured review, translation, and question answering over the current draft. Although prompts are customizable to accommodate different mathematical domains, the roles themselves are kept fixed to maintain stable interfaces and predictable behaviors.

LLM access is mediated through a lightweight streaming interface that relays requests to a user-specified endpoint without persistent storage of prompts or outputs. This design supports flexible deployment while keeping the annotation process transparent and controllable by human annotators.

B.4 Schema-constrained generation with review loop

To ensure machine-readable outputs, the generation process is constrained to produce records that can be deterministically parsed into the predefined schema. Generated drafts are automatically audited by a reviewer agent that checks for structural validity, answer consistency, and potential ambiguities.

When a draft fails review, a summarized set of issues is fed back to the generator and the process is repeated for a bounded number of rounds. If no fully satisfactory version is obtained within the allowed iterations, the latest draft is retained but explicitly marked as requiring further human inspection. This review loop improves robustness while preserving annotation throughput.

B.5 MathJax-aware normalization

Mathematical notation is normalized using a renderer-aware procedure. Instead of relying on generic LaTeX linting, the system detects rendering errors by typesetting snippets with MathJax and extracting parser-level error signals. A dedicated normalization agent then proposes minimal corrections that preserve semantics while resolving rendering failures. This mechanism is applied both interactively and in batch mode, improving consistency across the dataset.

B.6 Multimodal context and integrity checks

For problems involving visual information, images are attached as explicit multimodal inputs during generation and review. Image files are stored locally and referenced via relative paths in exported datasets. The system maintains consistency checks between records and referenced images, enabling safe renaming, re-importing, and cleanup operations.

B.7 Human-in-the-loop curation

Throughout the process, annotators remain in full control. All prompts and configurations are editable, generation histories are retained for inspection, and a dedicated question-answering assistant can be queried without modifying the current draft. Mandatory-field checks and explicit review status indicators reduce the risk of exporting incomplete or unreliable instances.

Overall, this system provides a lightweight yet structured framework for LLM-assisted dataset curation, balancing automation with expert oversight and enabling the construction of high-quality, reproducible benchmarks such as TopoEval.

Table 5: Overall accuracy (%) under three inference variants: Base (Instruct model), Base+CoT prompting, and Thinking variants. Missing entries are marked as "–".

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	–

Table 6: Overall accuracy (%) with and without 3-sample majority voting. Δ reports absolute changes in percentage points (pp) relative to the base setting of the same model.

Model line	Base	Base+3vote	Δ (pp)
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

C Additional Results

This section reports supplementary experimental results that are not included in the main paper due to space constraints. Specifically, Table 8 presents fine-grained performance across the 12 topology subfields for additional models beyond those shown in Table 2, providing a more complete view of model behavior. Tables 5 and 6 further analyze the effects of prompt-based inference strategies, including Chain-of-Thought prompting and multi-sample majority voting, focusing on their impact on overall accuracy. These results support the observations in the main text that such strategies lead to inconsistent and model-dependent performance gains under our experimental setting.

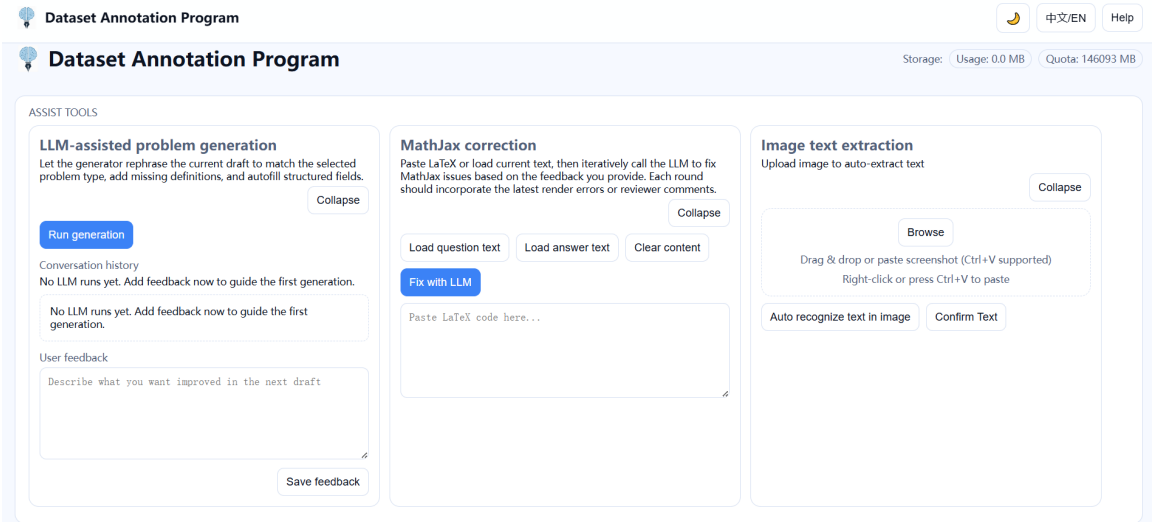


Figure 7: **Overview of the LLM-assisted annotation and adaptation system.** The interface supports structured problem editing, role-specialized LLM interactions, automatic review and normalization, and explicit human-in-the-loop control during dataset curation.

Table 7: **Abbreviation mapping for the 12 fine-grained topology subfields.** We also summarize each subfield's scope (including the major branch and typical concepts).

Abbr.	Full name	Scope (major branch + brief description)
TopSp	Topological Spaces	**Point-set.** Core axioms and basic notions of topological spaces (open/closed sets, bases, continuity).
Metric	Metric Spaces	**Point-set.** Metric structures, convergence and completeness, and links between metric and topological notions.
Knot	Knot Theory	**Geometric.** Knots/links, isotopy, invariants, and diagrammatic reasoning about entanglement.
Surf	Surface & MCG	**Geometric.** Classification and structures of surfaces; mapping class groups and basic surface invariants.
3M	3-Manifolds & Decompositions	**Geometric.** 3-manifold intuition and standard decompositions (e.g., splitting/cutting along surfaces), low-dimensional topology reasoning.
LDG	Low-Dimensional Geometry	**Geometric.** Geometric structures and intuition in low dimensions (e.g., planar/3D geometric-topological interactions).
Comb	Combinatorial Topology	**Geometric.** Piecewise-linear / combinatorial viewpoints (simplicial complexes, triangulations, discrete invariants).
Hom	Homology & Cohomology	**Algebraic.** Homology/cohomology as deformation-invariant descriptors (holes, cycles), basic computations and interpretations.
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 (e.g., flow/critical behavior).
Other	Other Topics	**Mixed.** A merged axis for rarer/overlapping subfields, including Compactness & Connectednes, continuous maps & homeomorphisms, topological constructions…

Table 8: **Additional results on the 12 fine-grained topology subfields.** We report accuracy (%) for the remaining models not shown in 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

D Prompt

This section summarizes the prompting protocols used in our experiments. We design distinct prompt templates for answering models and the judge model, reflecting their different functional roles in the evaluation pipeline. For answering models, we construct four standardized prompt variants. These variants cover both multiple-choice and free-form questions, and further distinguish whether explicit chain-of-thought (CoT) reasoning is requested. This design enables a controlled comparison of reasoning behaviors while keeping the task specification and output format consistent across models. All answering prompts enforce a strict JSON-only output format, which facilitates reliable automatic parsing and downstream evaluation. For the judge model, we adopt a single unified judging prompt. The judge model is used exclusively to extract final answers from model outputs and determine correctness with respect to the gold answers. It does not participate in problem solving and is never exposed to gold labels during generation. Using a unified judge prompt ensures consistent and reproducible evaluation across different models, question types, and inference settings.

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>
```

Judge Prompt

```
You are a strict evaluator. Decide whether the model answer should be counted as
    correct.
First, extract the final answer/result from the model answer (keep it short). Do
    NOT include any chain-of-thought.
Return ONLY a JSON object. No markdown, no extra text.
The JSON schema is:
{"verdict": "correct" | "incorrect" | "unjudgeable",
  "extracted_answer": string|null,
  "reason": string}
Example:{ "verdict": "correct", "extracted_answer": "3", "reason": "Matches the
    gold answer." }
Question:<QUESTION>
Gold Answer:<GOLD>
Model Answer:<MODEL_ANSWER>
```

E Case Study

This section presents a qualitative case study on a representative subset of problems from TopoEval. For each case, we analyze both the final answers and the underlying reasoning processes of different models. We categorize common failure modes into three main types: *knowledge deficiency*, where essential topological facts or definitions are missing or misapplied; *reasoning error*, where intermediate logical steps are inconsistent or internally contradictory; and *image understanding error*, where models fail to correctly perceive or track critical visual structures. Importantly, these case studies reveal that a correct final answer does not necessarily imply correct reasoning. In several instances, models arrive at the correct option through incomplete, fragile, or partially incorrect arguments.

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 $\chi(M) = 2 - 2g$. However, 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 (despite some redundant double-checking).

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 8: A differential topology problem together with representative model responses and error analysis. Correct answers and model's 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

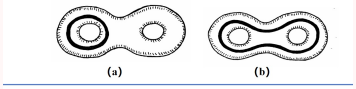
$$H_1(S;\mathbb{Q}) \cong \boxed{\mathbb{Q}}.$$

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

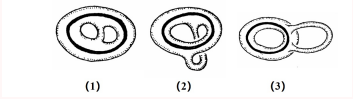
Figure 9: An algebraic topology problem on the Klein bottle together with representative model responses and error analysis. Correct answers and model's answers are highlighted in red.

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. Although local stretching and deformation are mentioned, the argument fails to maintain 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.

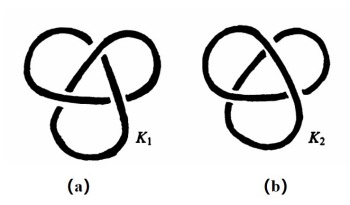
Evaluation: Correct.✓
Analysis: The global topological configuration is tracked consistently across all panels, and the intermediate deformation is used to justify the final ordering.

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 10: A topology deformation problem (pretzel with two holes) together with representative model responses and error analysis. Correct answers and model's 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, but the reasoning relies mainly on visual naming ("trefoil" vs. "untwistable") rather than 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) is misidentified as a figure-eight knot despite containing only three crossings. The actual over/under structure that permits simplification to the unknot is not examined.

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.

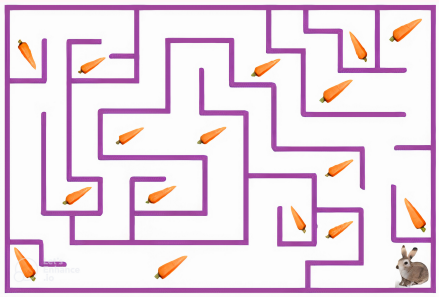
Evaluation: Correct.✓
Analysis: The classification is correct, but the argument appeals primarily to known knot types and invariants. A step-by-step Reidemeister simplification demonstrating the triviality of diagram (b) is not explicitly constructed.

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. However, after removing the crossing pair, a residual kink remains, 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 11: 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). Connectivity constraints imposed by the walls are not explicitly traced. Selects E.

Evaluation: Incorrect.✗
Error type: Image Understanding Error. The reasoning conflates *visible* carrots with *reachable* carrots. No explicit path tracing from the rabbit's starting position is performed, and isolated chambers are incorrectly assumed to be accessible.

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

Evaluation: Incorrect.✗
Error type: Reasoning Error. Although the model recognizes that wall constraints matter, it fails to maintain a stable global notion of reachability. Repeated recounting without a fixed connectivity map results in uncontrolled overcounting.

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 correctly distinguishes spatial regions but incorrectly assumes global connectivity. 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. It seems to be a hallucination: in the figure, moving left and then down from that corridor does not actually lead to a dead end with another carrot.

Figure 12: A maze-reachability problem with representative model responses and error analysis. Correct answers and model answers are highlighted in red.