DynaMath: A Dynamic Visual Benchmark for Evaluating Mathematical Reasoning Robustness of VLMs

ICLR 2025

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, Huan Zhang

Motivation

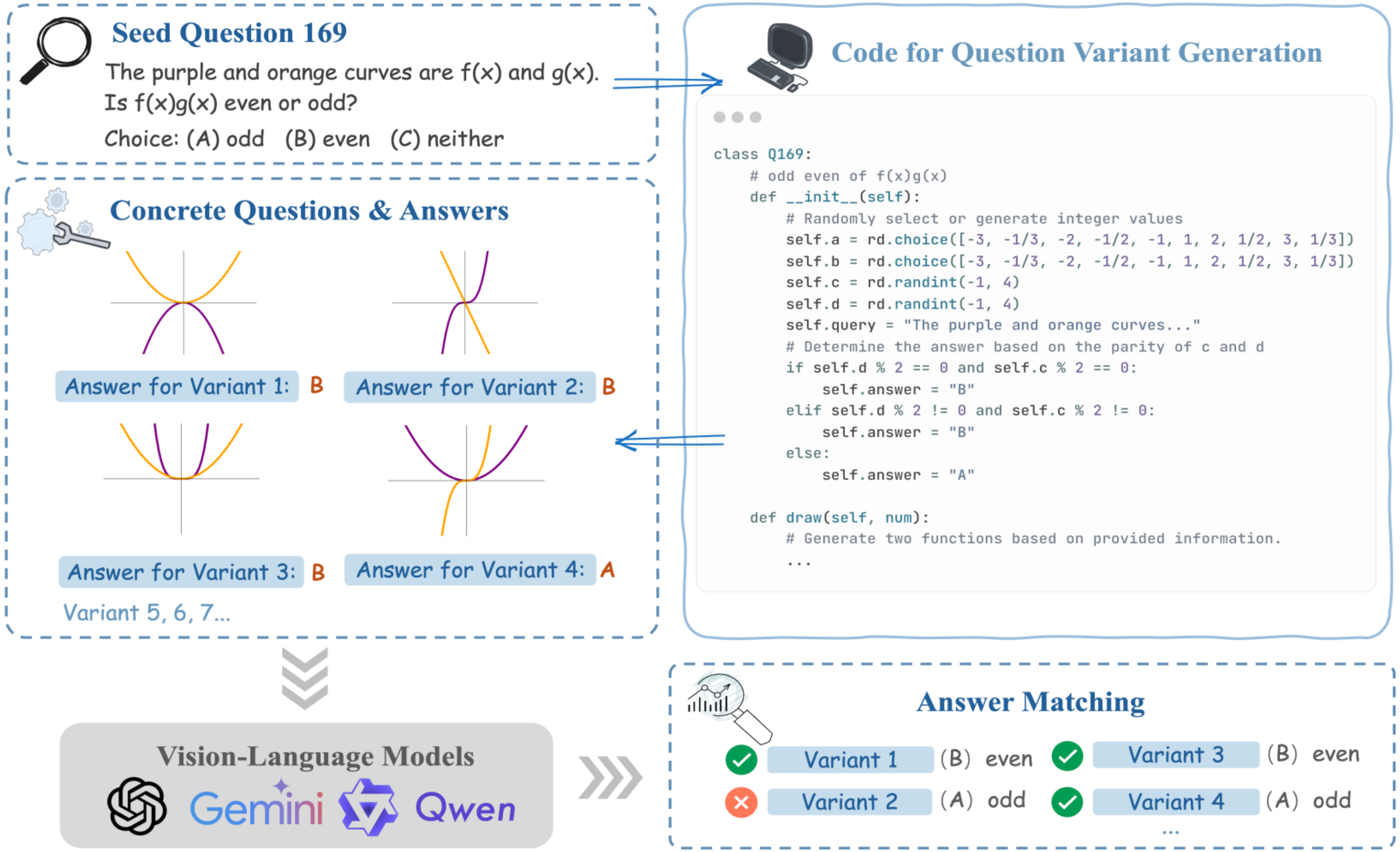
- VLMs have shown promise for mathematical reasoning tasks with visual contexts.
- Existing benchmarks are all static, limiting the assessment of robustness.
- **Key Challenge:** VLMs may struggle with variations of simple math problems that humans can solve easily.
- **Goal:** Design a benchmark to assess the robustness of VLMs under many variants of one seed question.

Contributions

- **DynaMATH Benchmark:** 501 seed questions generating 5010 variants
- Evaluated 14 state-of-the-art VLMs
- Unveils gaps between **average-case** and **worst-case accuracy**
- Provides insights into robustness and failure modes of VLMs.

Benchmark Design

- **Seed Questions:** curated from diverse math topics.
- **Program-Based Generation:** uses python programs to create variants.
- **Variants**: numerical values, geometric transformations, function types, etc.



Benchmark Design

Statistic	Number
Total *seed* questions (programs)	501
- Created from existing dataset	227 (45.3%)
- Newly designed questions	274 (54.7%)
Topics	
- Solid geometry (SG)	15 (3.0%)
- Puzzle test (PT)	17 (3.4%)
- Arithmetic (AR)	26 (5.2%)
- Scientific figure (SF)	45 (9.0%)
- Graph theory (GT)	48 (9.6%)
- Algebra (AL)	51 (10.2%)
- Plane geometry (PG)	77 (15.4%)
- Analytic geometry (AG)	97 (19.4%)
- Statistics (ST)	125 (25.0%)
Levels	
- Elementary school (EL)	63 (12.6%)
- High school (HI)	277 (55.3%)
- Undergraduate (UN)	161 (32.1%)
Question Types	
- Numerical questions	296 (59.1%)
- Multiple-choice questions	174 (34.7%)
- Free-form questions	31 (6.2%)

Table 1: Statistics of DYNAMATH.

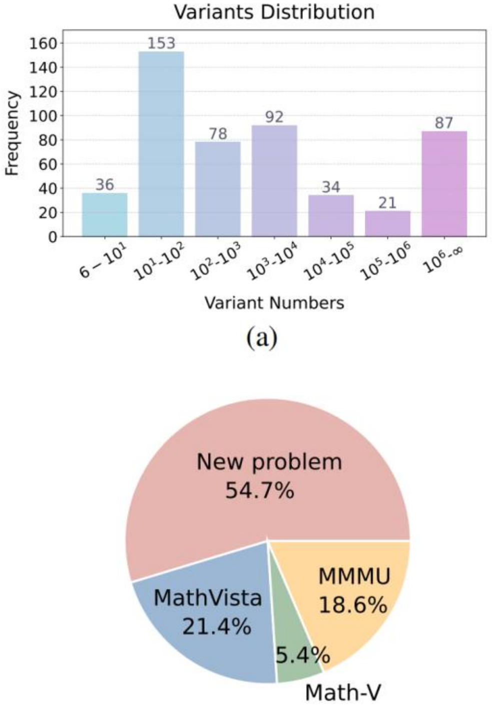
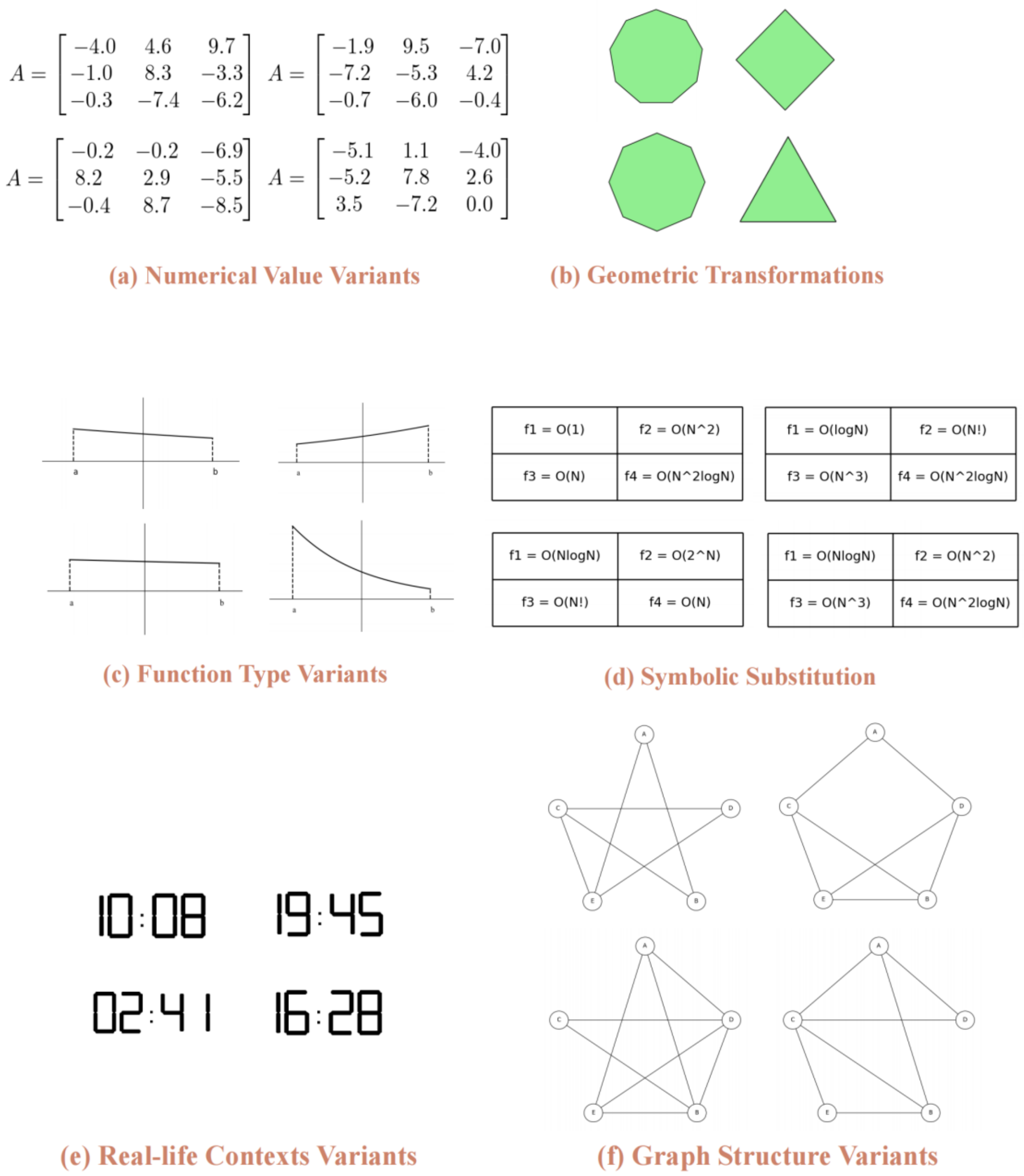
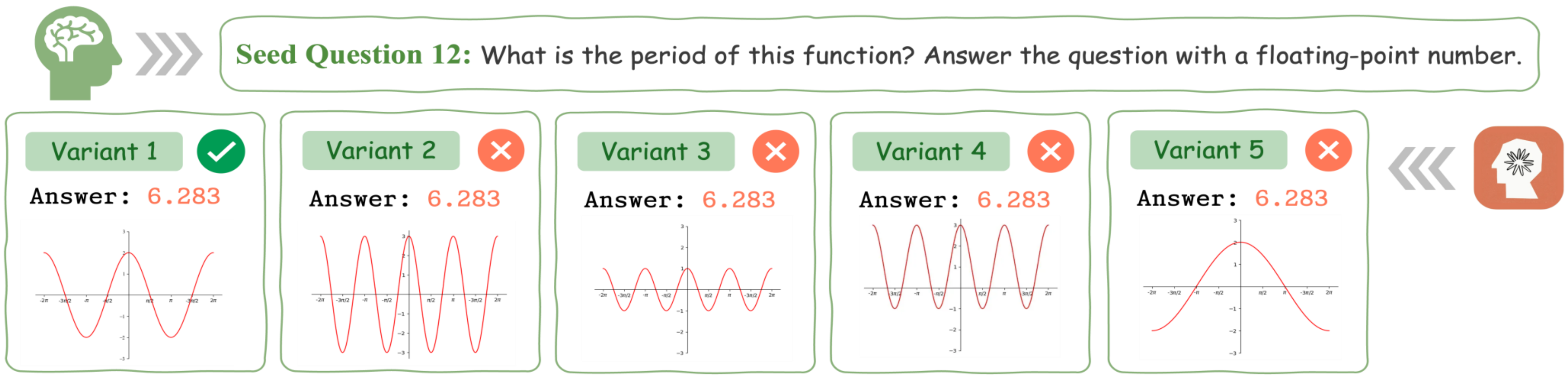


Figure 4: (a) Variant number distribution and (b) source composition of DYNAMATH.



Benchmark Problem Example I



Evaluation Metrics

- **Average-Case Accuracy (Aavg):** measures mean performance.

$$\mathcal{A}_{avg} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M}\sum_{j=1}^{M}\mathbb{I}[\mathrm{Ans}(i,j) = \mathbf{GT}(i,j)]$$

- **Worst-Case Accuracy (Awst):** Focuses on hardest variants.

$$\mathcal{A}_{wst} = \frac{1}{N}\sum_{i=1}^{N}\min_{j\in[1,M]}\mathbb{I}[\mathrm{Ans}(i,j) = \mathbf{GT}(i,j)]$$

- **Reasoning Robustness (RR):** ratio of Aswt to Aavg.

$$RR = \frac{\mathcal{A}_{wst}}{\mathcal{A}_{avg}},$$

Evaluation Results

Model	ALL	PG	SG	AG	AL	PT	GT	ST	SF	AR	EL	HI	UN
Closed-sourced Large Multimodal Models (LMMs)													
Zero-shot GPT-4o	63.7	56.8	52.0	61.0	76.9	51.8	58.1	69.3	62.4	61.5	68.6	61.8	36.8
Zero-shot Claude-3.5	64.8	49.9	49.3	55.3	81.0	44.1	69.4	78.2	62.2	61.2	66.7	62.6	33.3
Zero-shot Gemini Pro 1.5	60.5	52.7	42.7	61.6	70.8	20.6	65.2	69.8	50.2	54.2	62.9	59.2	37.1

Average Accuracy

Model	ALL	PG	SG	AG	AL	PT	GT	ST	SF	AR	EL	HI	UN
Closed-sourced Large Multimodal Models (LMMs)													
Zero-shot GPT-4o	34.7	37.7	33.3	25.8	54.9	11.8	18.8	38.4	35.6	46.2	46.0	34.3	31.1
Zero-shot Claude-3.5	35.3	22.1	26.7	18.6	62.7	23.5	27.1	53.6	24.4	42.3	49.2	33.2	33.5
Zero-shot Gemini Pro 1.5	26.9	28.6	20.0	19.6	39.2	5.9	22.9	35.2	15.6	30.8	41.3	26.7	21.7

Worst-case Accuracy

Evaluation Results

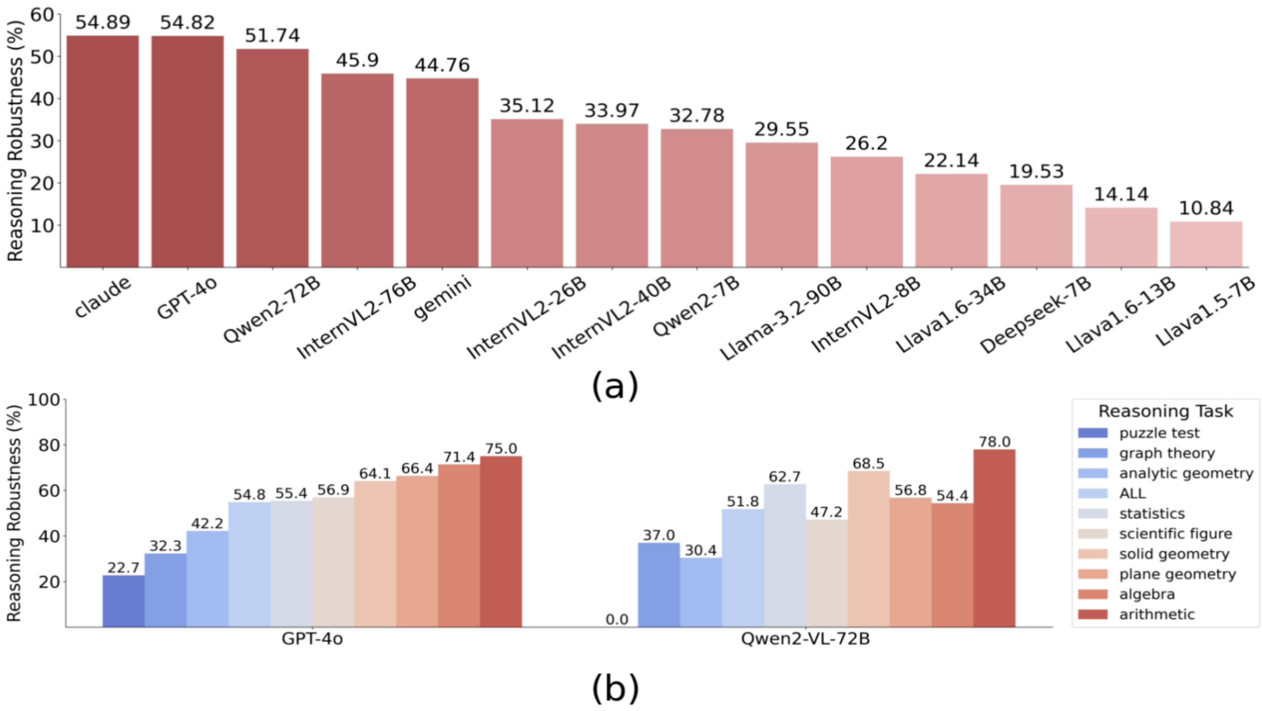


Figure 5: Comparing reasoning robustness across different (a) models and (b) topics.

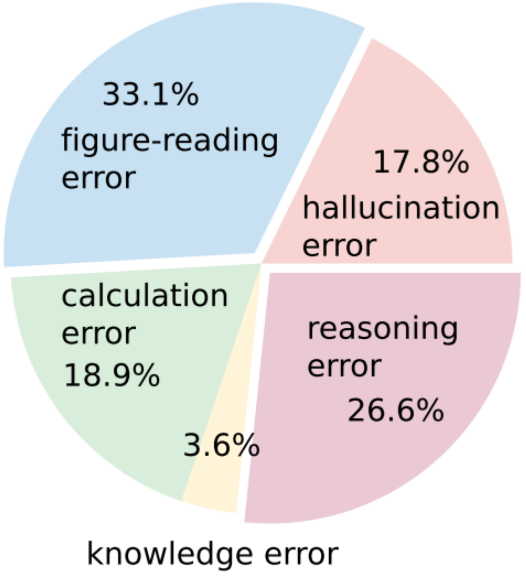


Fig: Failure Analysis

Conclusion and Future Work

- **Conclusion:**
  - Significant robustness gaps in VLMs.
  - Dynamic benchmark provides a controllable way to examine model's robustness.
- **Future Work:**
  - Scale up the benchmark design with the aid of LLMs.
  - Improve model's reasoning robustness.