Math Vision: Fine-tuning Multi-modal LLMs for Geometry Problem Solving

Atul Shreewastav	**(THA077BCT013)**
Bidhan Acharya	**(THA077BCT015)**
Nischal Paudel	**(THA077BCT028)**
Yugratna Humagain	**(THA077BCT047)**

Under supervision of
Er. Bibat Thokar and Er. Sushil Dyopala

Department of Electronics and Computer Engineering
Institute of Engineering, Thapathali Campus

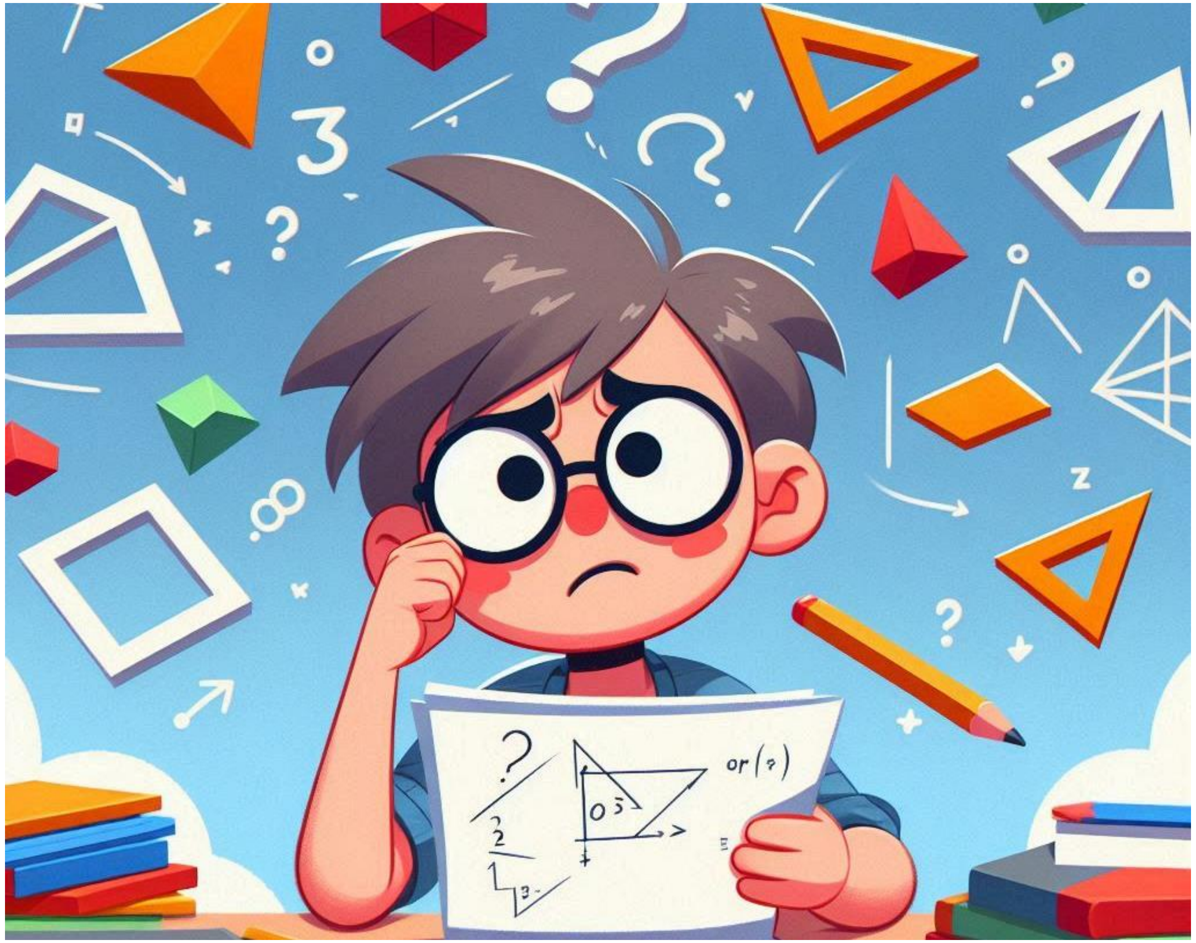
July 22, 2024

Outlines

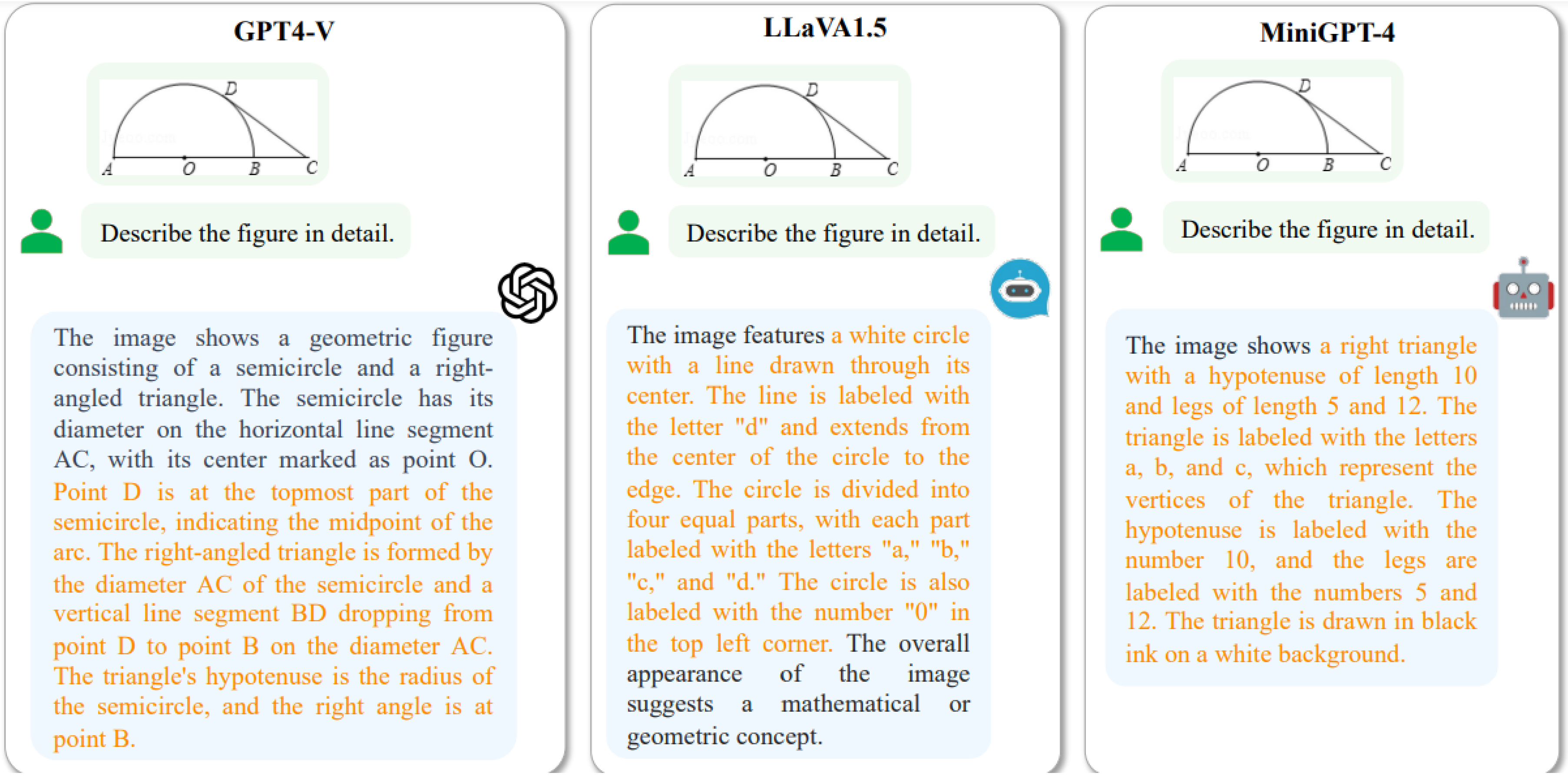
- Motivation
- Objectives
- Scope of Project
- Project Applications
- Methodology
- Dataset Exploration
- Results
- Analysis and Discussion
- Remanning Tasks
- References

Motivation

- Existing LLMs struggle with geometric shapes.
- High incidence of hallucinations and inaccuracies.
- Geometry problem solving is essential for education, engineering, and architecture



Problem with existing LLMs



Objectives

- Fine-tune a Multi-modal Large Language Model (MLLM) to solve geometric problems.
- Enhance accuracy in geometric problem-solving.
- Reduce hallucinations and irrelevant results.
- Provide precise, step-by-step solutions.

Scope of Project

Project Capabilities:

- Interpret and solve geometry problems from text and images.
- Providing detailed solutions in text and mathematical expressions.
- Minimizing model hallucinations for accurate problem interpretation.

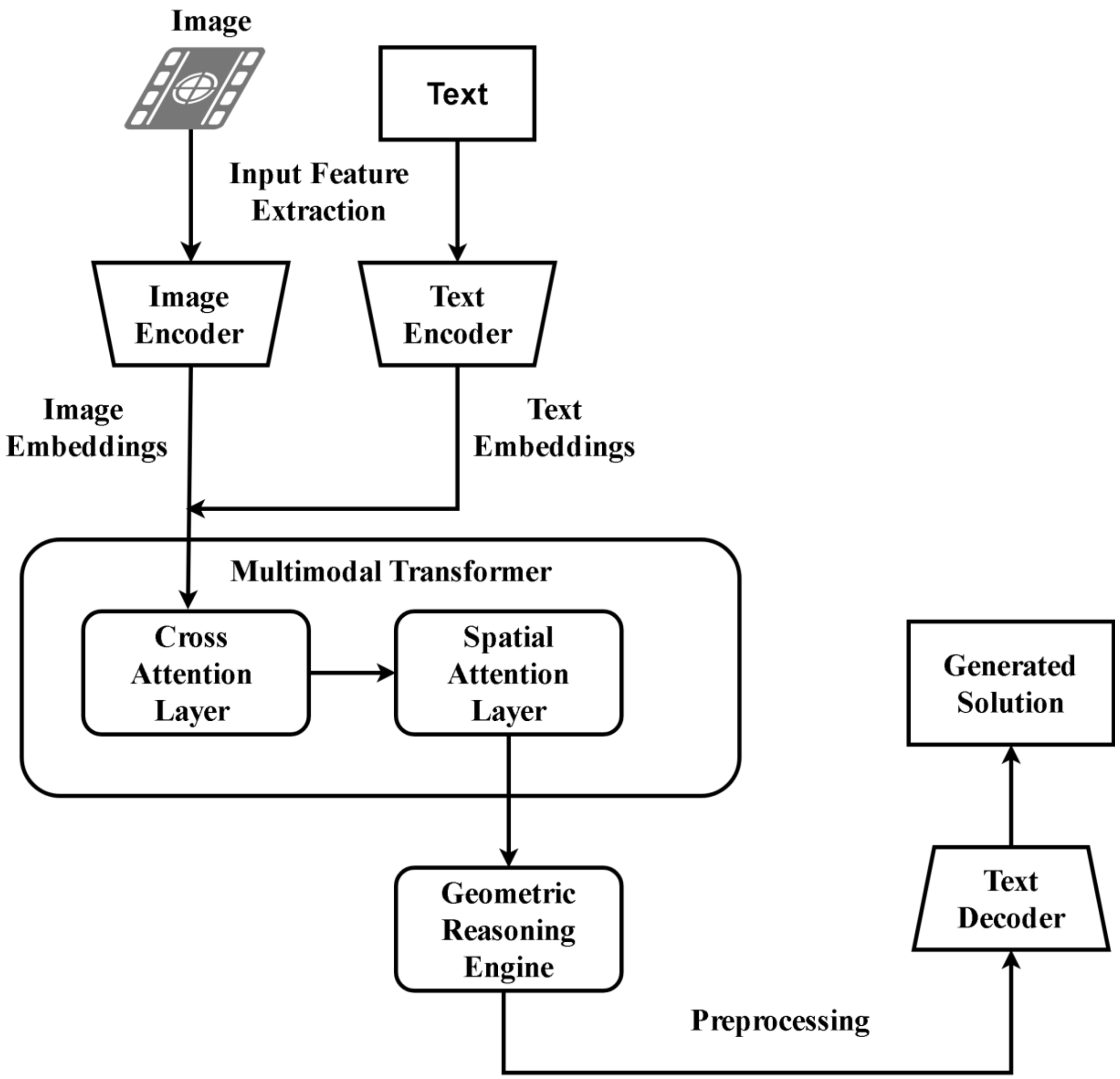
Project limitations:

- Specialized in geometry, not addressing other math fields.
- Model effectiveness depends on quality and diversity of training datasets.
- Regular updates and maintenance needed to incorporate new geometric theories and techniques.

Project Application

- Education:
  - Intelligent tutoring systems, personalized learning.
- Engineering and Architecture:
  - Automatic geometric modeling, error detection
- Scientific Research:
  - Automatic geometric data analysis
- Robotics and Computer Vision:
  - Enhanced path planning, object recognition.

Methodology-[1] (System Architecture)



Methodology-[2]
(System Architecture)

- **Text Encoder:**
  - Converts text into dense vector representation using models like BERT or GPT-3.
  - Captures semantic meaning of textual problem descriptions.
- **Image Encoder:**
  - Uses CNNs like RestNet to extract visual features from geometric diagrams.
  - Transforms visual data into compact representations.

Methodology-[3]
(System Architecture)

- **Multimodal Transformer:**
  - Cross-Attention Layer: Aligns and integrates textual data and visual information.
  - Spatial Attention Layer: Focuses on specific image regions relevant to the problem.
- **Geometric Reasoning Engine:**
  - Applies geometric algorithms using libraries like SymPy for calculation.
- **Text decoder:**
  - Generates coherent, step-by-step solutions in natural languages.

Methodology-[4]
(Input Types)

- **Vision Intensive:**
  - Primarily visual data with minimal text.
- **Vision Dominant:**
  - Significant visual data with moderate text.
- **Text Lite:**
  - Primarily text with some or no visual data.
- **Text Dominant:**
  - Mainly textual data with minimal visual input.
- **Vision Only:**
  - Solely visual data without text.

Methodology-[5]
(Text Preprocessing)

- **Tokenization:**
  - Break text into tokens (words/sub words)
  - Utilizes algorithm like BERT tokenizer
- **Padding and Truncation:**
  - Ensures uniform sequence length
  - Implements techniques to manage varying text lengths effectively

Methodology-[6]
(Image Preprocessing)

- **Resizing:**
  - Adjusts image dimensions for consistency
  - Techniques include bilinear interpolation
- **Normalization:**
  - Scales pixel values to a standard range
  - Algorithms such as min-max normalization are applied

Methodology-[7]
(Text Feature Extraction)

- **Token Embedding:**
  - Converts words into numerical vectors.
  - Represents each word uniquely in the context.
- **Contextual Embedding:**
  - Captures meaning based on surrounding text.

Methodology-[8]
(Image Feature Extraction)

- **Hierarchical Features:**
  - Extracts patterns from low-level to high-level**.**
  - Identifies shapes, edges, and complex structures.
- **Spatial Features:**
  - Analyzes relationships between elements.
  - Identifies positional and structural data.

Methodology-[9]
(Multimodal Alignment Techniques)

- **Cross-Modal Attention:**
  - Integrates textual and visual information
  - Utilizes algorithms such as Cross-Modal Transformers (CMT)
- **Joint Embedding:**
  - Maps text and image features into a common space
  - Algorithms like Dual-Stream Networks facilitate alignment
- **Spatial Attention:**
  - Focuses on relevant spatial regions in images
  - Enhances alignment by attending to spatial relationships

Methodology-[10]
(Feature Fusion)

- **Concatenation:**
  - Merges text and images features sequentially.
  - Enables joint representation learning
- **Attention Fusion:**
  - Focus on relevant features across modalities
  - Utilizes algorithms such as Multimodal Attention Mechanisms

Methodology-[11] (OpenBMB MiniCPM-V-2)

- **Architecture:**
  - **Base Model:** Transformer-based architecture
  - **Multimodal Input:** Handles both text and image inputs
  - **Parameters:** Approximately 2.6 billion parameters
  - **Layers:** Stacked transformer layers with attention mechanisms

Methodology-[12] (MILVLG/imp-v1-3b)

- **Architecture:**
  - **Base Model:** Transformer-based architecture
  - **Multimodal Input:** Handles both text and image inputs
  - **Parameters:** 3 billion parameters
  - **Layers:** Multiple transformer layers with self-attention mechanisms
  - **Visual Encoder:** Extracts image features for integration with text

Methodology-[13] (LLava-HF/LLava-1.5-7B-HF)

- **Architecture:**
  - **Base Model:** GPT-3.5 architecture
  - **Multimodal Input:** Combines text and images inputs
  - **Parameters:** 7 billion parameters
  - **Vision Component:** Pre-trained vision encoder (CLIP) for image processing
  - **Text Component:** Large language model for natural language understanding
  - **Fusion Mechanism:** Cross-attention layers to integrate visual and textual information.

Dataset Exploration-[1]

- **Total Records:** 600 Vision Dominant problems.
- **Dataset Structure:**
    - **Question:** Contains text and associated image.
    - **Choices:** Multiple-choice options for answers.
    - **Answer:** Correct answer.
    - **Annotations:** Metadata about the problem (type, figure, theorem, difficulty level).
    - **Solution:** Step-by-step solution breakdown.

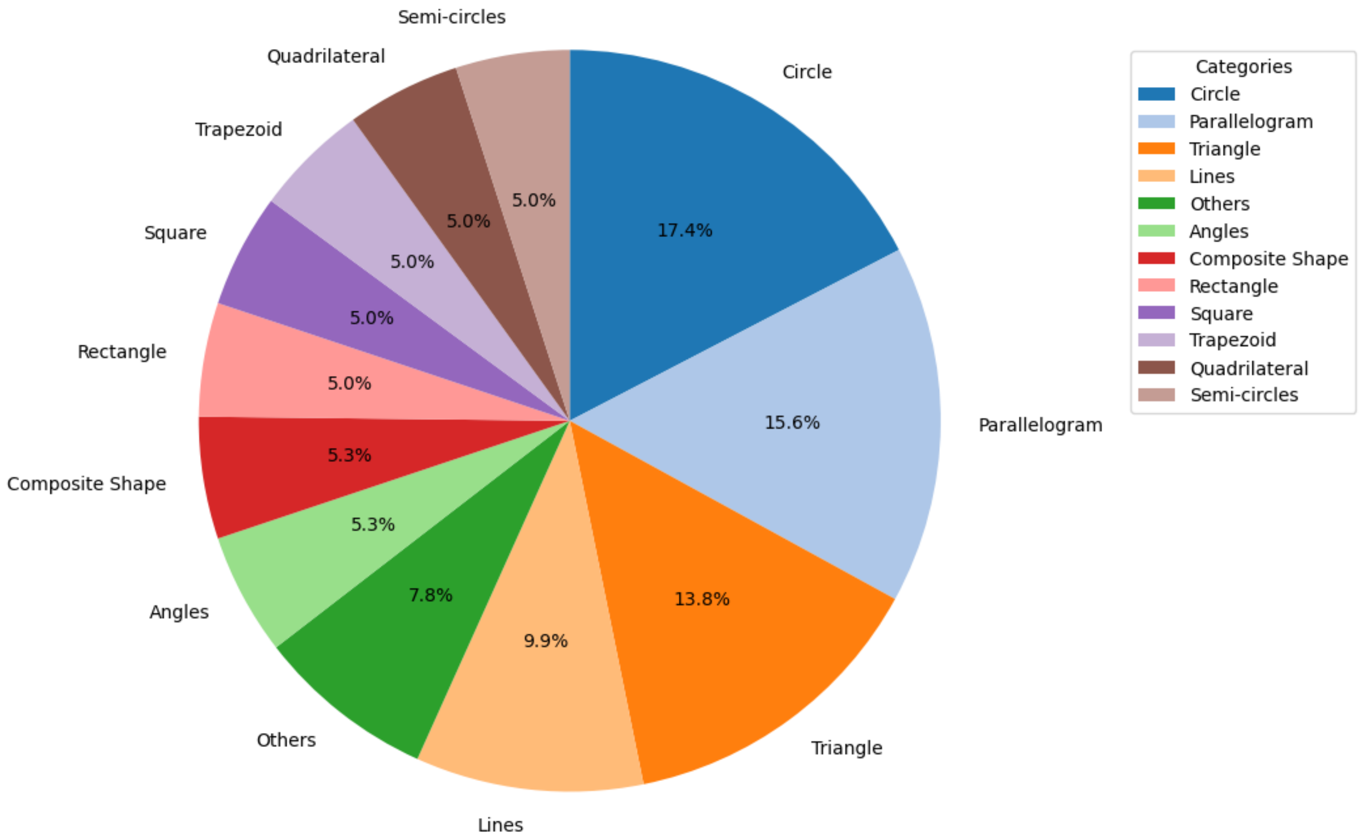
Dataset Exploration-[2] (Dataset Snippet)


```
[
    {
        "id": "549",
        "question": {
            "text": "Find the length of ⌒JK. Round to the nearest hundredth.",
            "image_url": "/1.png"
        },
        "choices": ["0.52", "0.79", "1.05", "1.57"],
        "answer": "1.05",
        "annotations": {
            "type": "Vision Dominant",
            "figure": "Circle",
            "theorem": "Arc Length",
            "difficulty level": "medium"
        },
        "solution": {
            "steps": [
                {
                    "step": 1,
                    "description": "Identify the given radius and central angle of the circle."
                },
                {
                    "step": 2,
                    "description": "Use the formula for arc length: L = rθ."
                },
                {
                    "step": 3,
                    "description": "Convert the central angle from degrees to radians: θ = 30° *
                     (π/180) = π/6."
                },
                {
                    "step": 4,
                    "description": "Substitute the radius and angle into the formula: L = 2 * (π/6)."
                },
                {
                    "step": 5,
                    "description": "Calculate the arc length: L = π/3 ≈ 1.05."
                }
            ],
            "final_answer": {
                "description": "The length of ⌒JK is approximately 1.05."
            }
        }
    }
]
```

Dataset Exploration-[3] (Data Annotation and Metadata)

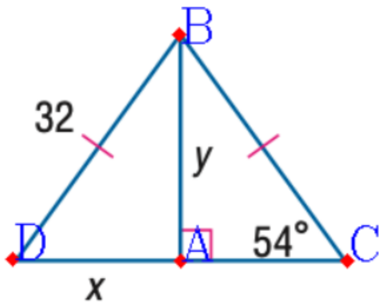
- **Annotations:**
  - **Types:** Vision Intensive, Vision Dominant, Text Lite, Text Dominant, Vision Only
  - **Figure:** Type of geometric figures (triangle, circle, polygon)
  - **Theorem:** Relevant Geometric theorem used.
  - **Difficulty Level:** Easy, Medium, Hard

Dataset Exploration-[4] (Data Distribution)

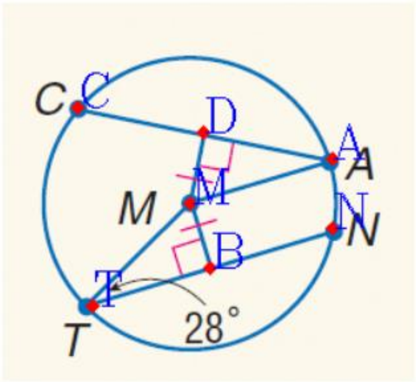


Result-[1] (OpenBMB MiniCPM-V-2)

Q2. Find x. Round to the nearest tenth.



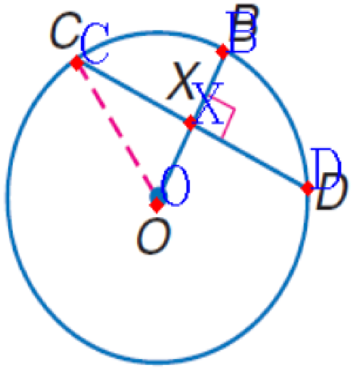
Q3. Find m∠CAM.




```
Answer for Question 2:
Models Output
The predicted answer is: 16
Actual Output:
The expected answer is: x is approximately 18.8.
------------------------------------------------------------------------------------------
Answer for Question 3:
Models Output
The predicted answer is: 28°
Actual Output:
The expected answer is: m∠CAM is 28°.
------------------------------------------------------------------------------------------
```

Result-[2] (MILVLG/imp-v1-3b)

Q1. Circle O has a radius of 13 inches. Radius OB is perpendicular to chord CD which is 24 inches long. Find OX.



Actual Ouptut:
60
Models Output:
Solution:

We can use the Pythagorean theorem to find the length of OX.

OX^2 = OB^2 + CD^2

OX^2 = (13)^2 + (24)^2

OX^2 = 169 + 576

OX^2 = 745

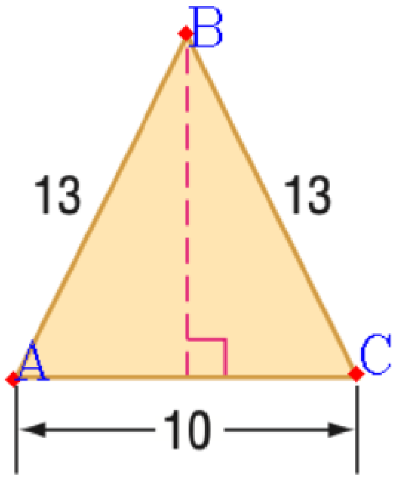
OX = sqrt(745)

OX = 26.83 inches (rounded to two decimal places)

Therefore, the length of OX is approximately 26.83 inches.

Result-[3]
(LLava-HF/LLava-1.5-7B-HF)

Q. Find the area of the figure.



Actual Output
Steps:
Step 1: Identify the lengths of the sides of the triangle: two equal sides of 13 units each and a base of 10 units.
Step 2: Determine the height of the triangle by dividing it into two right-angled triangles. Each right triangle will have a hypotenuse of 13 units and a base of 5 units (half of 10 units).
Step 3: Use the Pythagorean theorem to find the height (h) of the triangle. The Pythagorean theorem states that in a right-angled triangle, $a^2 + b^2 = c^2$, where c is the hypotenuse.
Step 4: Set up the equation: $5^2 + h^2 = 13^2$. This simplifies to $25 + h^2 = 169$.
Step 5: Solve for h: $h^2 = 169 - 25 = 144$, so $h = \sqrt{144} = 12$ units.
Step 6: Calculate the area of the triangle using the formula: Area = 0.5 * base * height. Substitute the known value s: Area = 0.5 * 10 * 12.
Step 7: The area of the triangle is 60 square units.
Final Answer: The area of the triangle is 60 square units.
****

Models Output
USER:
Find the area of the figure.
ASSISTANT: To find the area of the figure, we need to determine the dimensions of the triangle. The triangle has a ba se of 10 units and a height of 13 units. The area of the triangle can be calculated using the formula:

Area = (Base * Height) / 2

Area = (10 * 13) / 2

Area = 130 / 2

Area = 65

So, the area of the triangle is 65 square units.

Discussion and Analysis –[1] (Model Performance Comparison)

- **LLaVA-1.5-7B-HF:** Most promising results, excellent integration of visual and textual data, effective use of CLIP and cross-attention layers. (Accuracy = 11.17%)
- **OpenBMB MiniCPM-V-2:** Competent but lagged in accuracy and step-by-step solution. (Accuracy = 7.13%)
- **MILVLG/imp-v1-3b:** Adequate but struggled with complex visual data. (Accuracy = 4.96%)

Discussion and Analysis- [2] Challenges Identified

- **Model Hallucinations:** Significant issue across models; LLaVA had fewer instances but needs further improvement.
- **Visual Data Handling:** Varied effectiveness in interpreting and utilizing visual data; accurate feature extraction and integration are critical for success.

Remaining Tasks

- Expand Dataset to Diverse Modalities
  - Vision Intensive, Text Lite, Text Dominant, Vision Only
  - Current Progress: Built 600 records for Vision Dominant (significant visual data with moderate text).
- Evaluate and fine-tunning the best model among:
  - LLaVA-1.5-7b-h, MILVLG/imp-v1-3b and openbmb/MiniCPM-V-2
  - Focus primarily on LLaVA-1.5-7b-h for optimal performance.
- Create an User Interface

References-[1]

[1] J. Chen, J. Tang, J. Qin, et al., “Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning,” in Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1- 6, 2021, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., ser. Findings of ACL, ACL/IJCNLP 2021, Association for Computational Linguistics, 2021, 513–523. DOI: 10.18653/v1/2021.findings-acl.46.

[2] Y. Z. H. L. Z. G. P. Q. A. Z. P. L. K.-W. C. P. G.H. L. Renrui Zhang, Dongzhi Jiang, “Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?” in arXiv, 2024

References-[2]

[3] M.-L. Zhang, F. Yin, Y.-H. Hao, and C.-L. Liu, “Plane geometry diagram parsing,” arXiv preprint arXiv:2205.09363, 2022.

[4] Y. Hao, M. Zhang, F. Yin, and L.-L. Huang, “Pgdp5k: A diagram parsing dataset for plane geometry problems,” in 2022 26th International Conference on Pattern Recognition (ICPR). IEEE, 2022, pp. 1763–1769.

[5] M.-L. Zhang, F. Yin, and C.-L. Liu, “A multi-modal neural geometric solver with textual clauses parsed from diagram,” arXiv preprint arXiv:2302.11097, 2023.

THANK YOU