

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

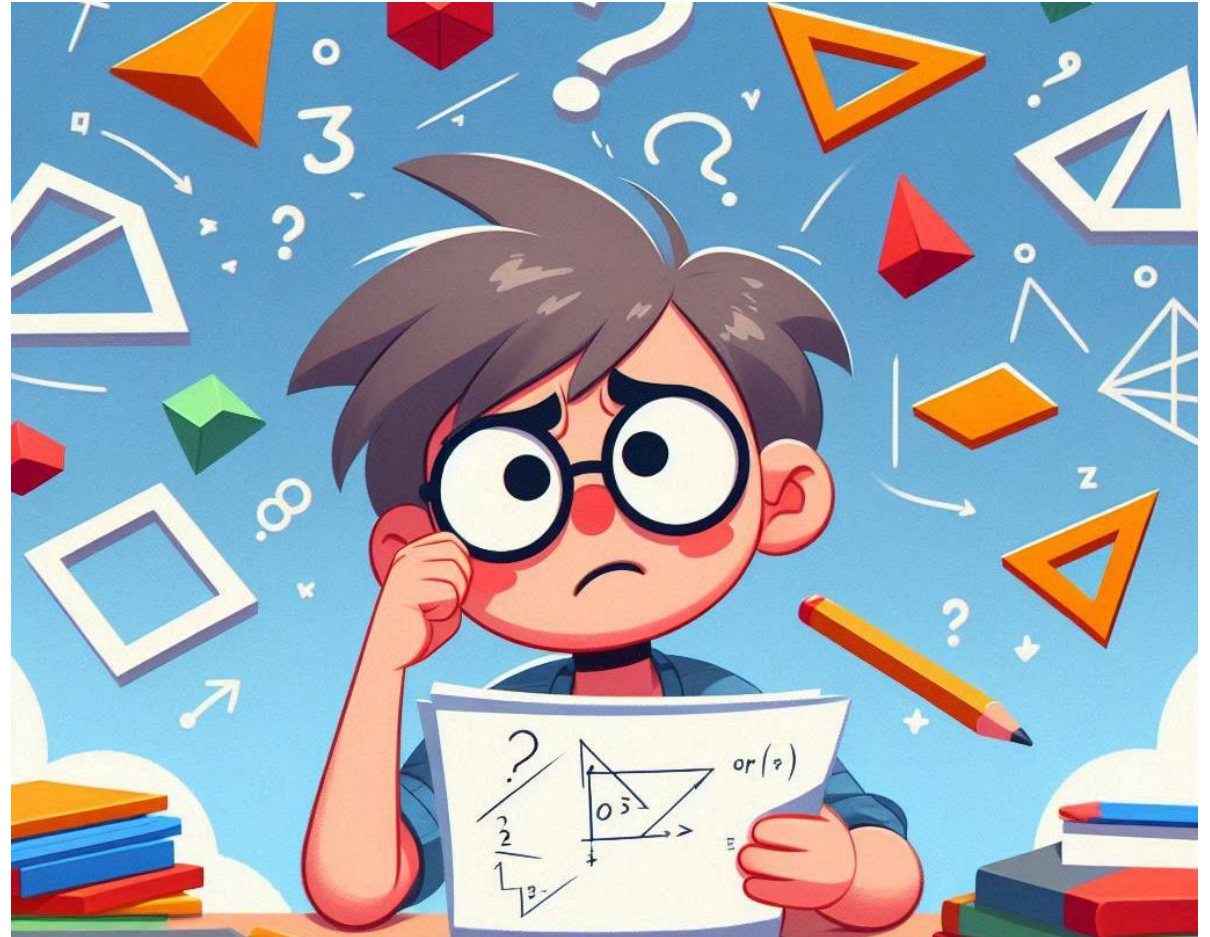
August 9, 2024

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



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.

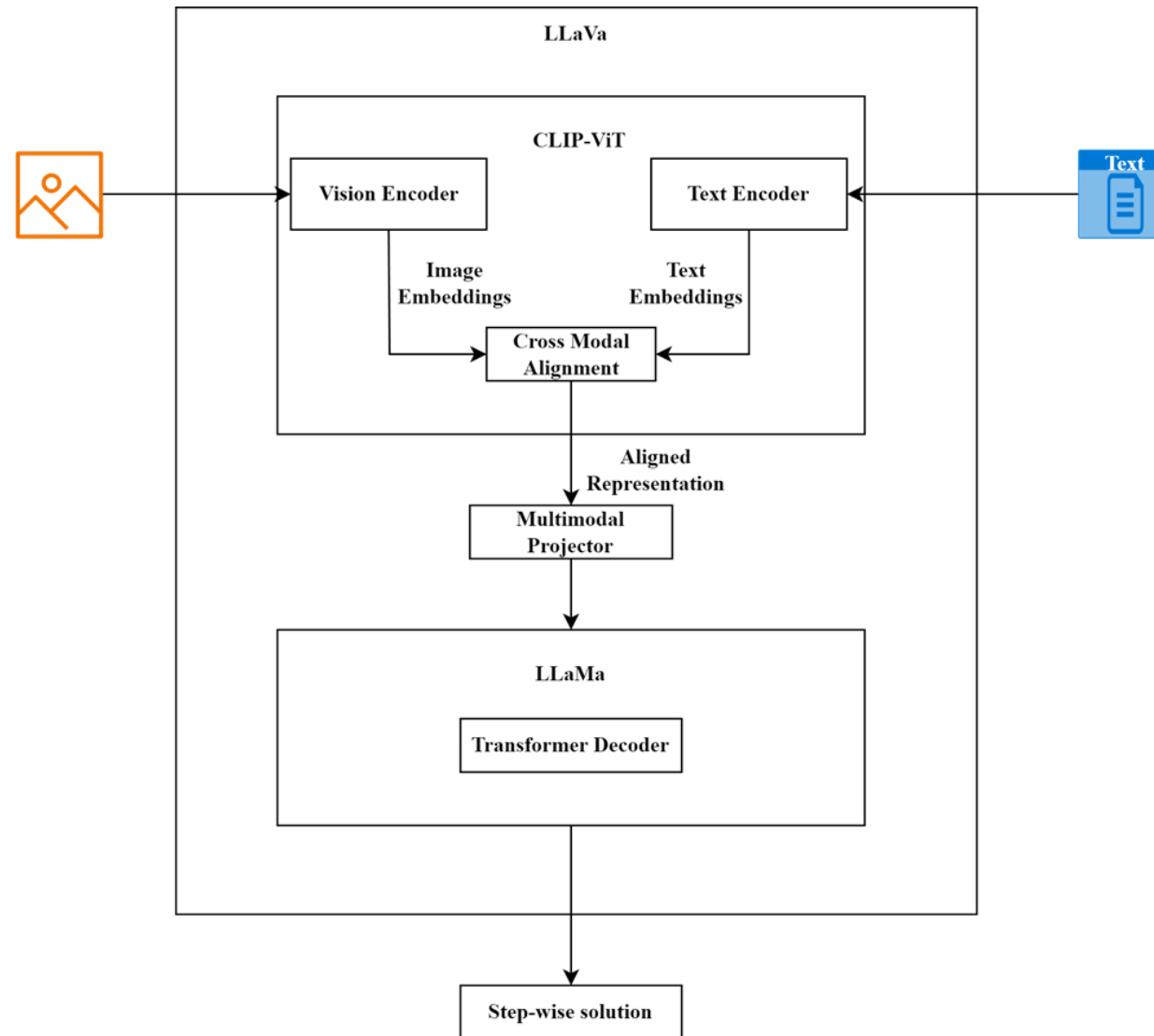
Project limitations:

- 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)

- **Inputs:** System takes two types of inputs
- **CLIP-ViT:** CLIP-ViT (Contrastive Language-Image Pre-training Vision Transformer) module is responsible for processing both visual and textual inputs.
 - **Vision Encoder:** This component processes the image input and generates image embeddings.
 - **Text Encoder:** This component processes the text input and generates text embeddings.

Methodology-[3]

(System Architecture)

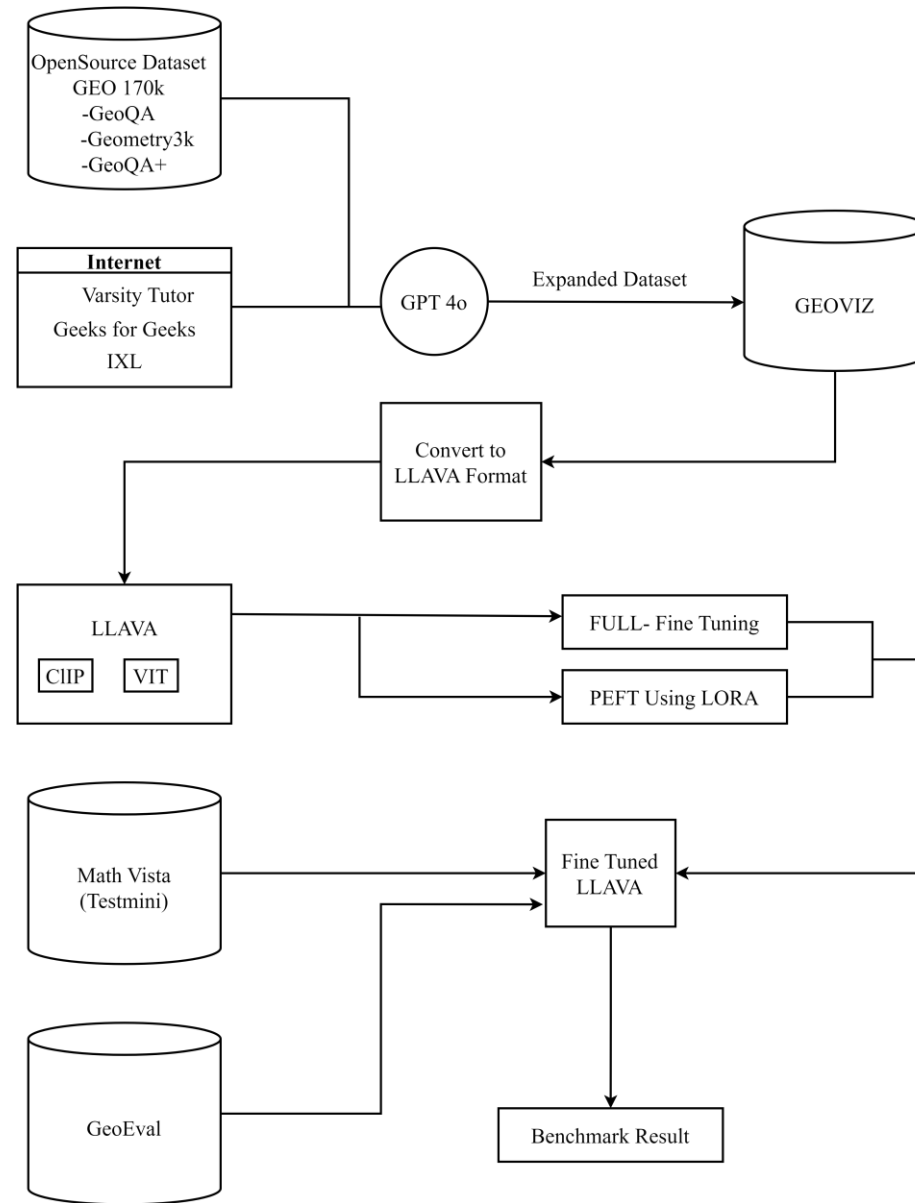
- **Cross Modal Alignment:**
 - The image embeddings and text embeddings are aligned.
 - Representations from both modalities are compatible and can be combined meaningfully.
- **Multimodal Projection**
 - Aligned representations from cross-modal alignment are projected into a common multimodal space.
 - Ensures the combined representation retains important features from both visual and textual data.

Methodology-[4]

(System Architecture)

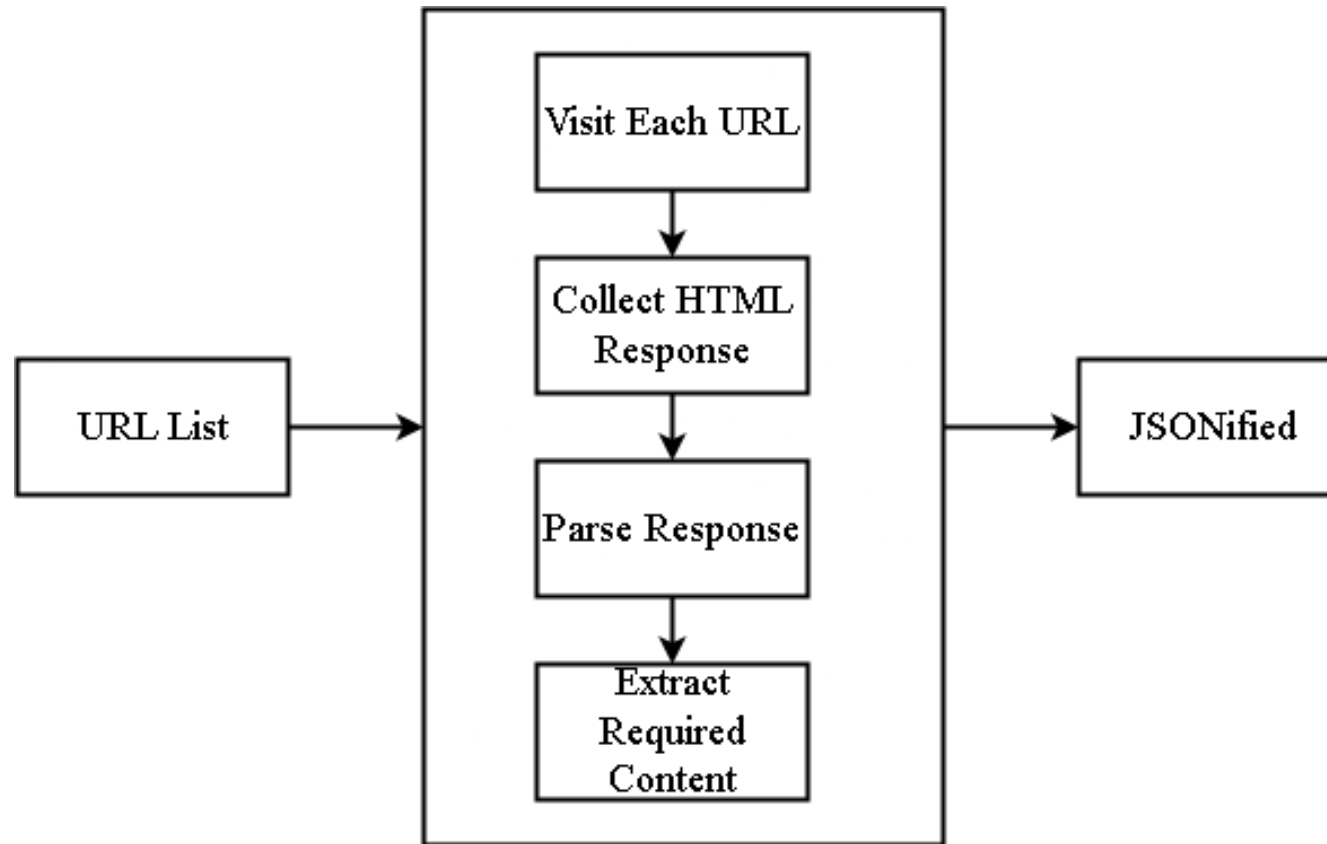
- **LLAMA**
 - **Aligned Representation Input:** LLaMa uses the aligned representation from the multimodal projected as input
 - **Transformer Decoder:** Processes the multimodal representation and generates the final output

Methodology-[5] (Implementation)



Methodology-[6]

(Text Datasets Preparation)



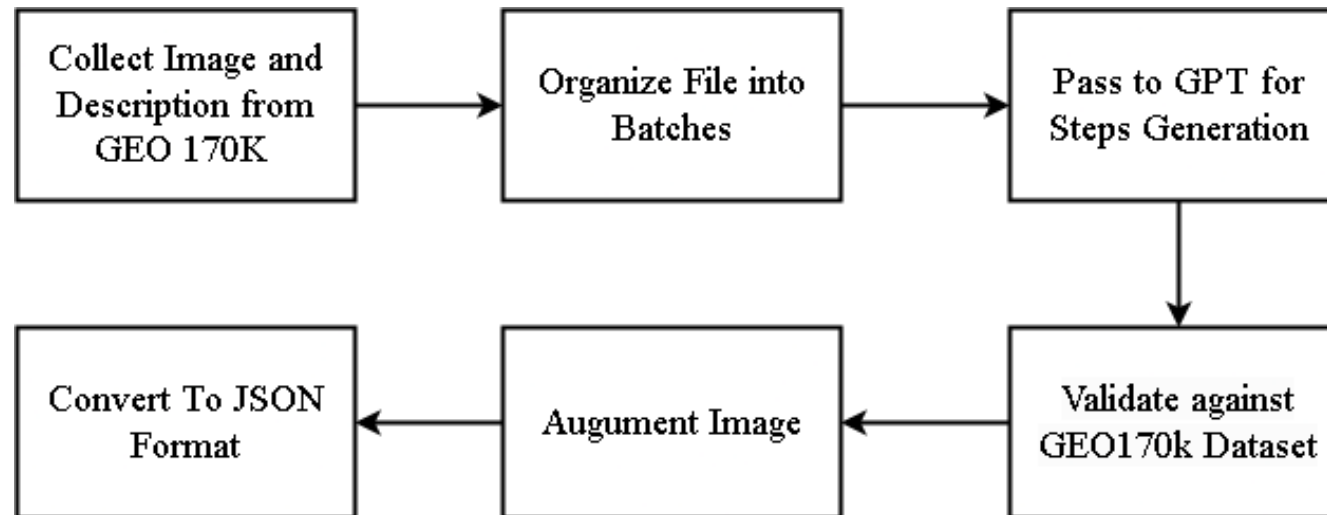
Methodology-[7]

(Text Datasets Preparation)

- Visit each URL for web scraping to collect HTML responses.
- Collect and store HTML responses for further processing.
- Parse HTML responses to extract relevant content.
- Extract required content, including problem statements, answers, and explanations.
- Convert extracted content into JSON format for easy integration and manipulation in the model training pipeline.

Methodology-[8]

(Image Dataset Preparation)



Methodology-[9]

(Image Datasets Preparation)

- Collecting diverse images and descriptions from the GEO 170K dataset for training.
- Organizing the data into manageable batches for efficient handling and systematic processing.
- Passing the organized data to GPT – 4o model for step-by-step solutions.
- Augmenting the images using techniques like rotation, scaling, color jittering, perspective transformations, and grayscale conversion.
- Converting the data into JSON format for easy integration and retrieval during model training and evaluation.

Methodology-[10]

(Data Augmentation)

- Affine Transformations:
 - Rotation
 - Scaling
 - Translation
 - Shearing
- GrayScale Conversion
- Color Jittering (Brightness, Contrast, Saturation, Hue Adjustment)
- Perspective Transformation

Dataset Exploration-[1]

- **GeoViz Dataset:** 1000 Vision problems and 1000 text data.
- **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]

(Text Data Annotation and Metadata)

- **Annotations:**
 - **Shape:** Type of geometric figures (triangle, circle, polygon)
 - **Theorem:** Relevant Geometric theorem used.
 - **Difficulty Level:** Easy, Medium, Hard

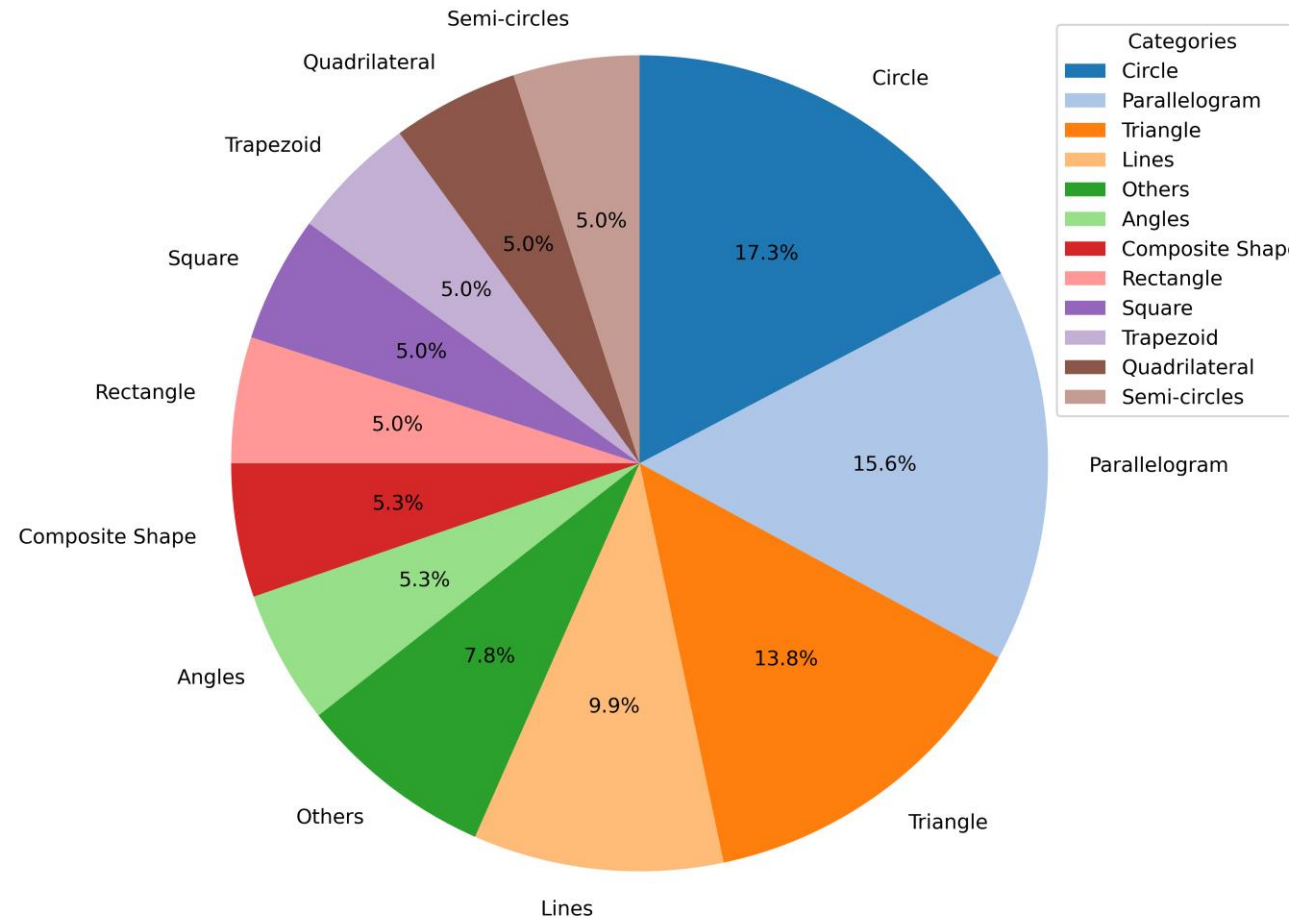
Dataset Exploration-[3] (Text Dataset Snippet)

8/9/2024

```
{
  "id": 53,
  "question": {
    "text": "Find the length of the chord of a circle with radius 10 units, if the chord is 12 units long."
  },
  "choices": [
    "8 units",
    "10 units",
    "12 units",
    "6 units"
  ],
  "answer": "8 units",
  "annotations": {
    "shape": "Circle",
    "theorem": "Chord Length Formula",
    "difficulty level": "Hard"
  },
  "solution": {
    "steps": [
      {
        "step": 1,
        "description": "The length of the chord can be found using the formula:  $d = \sqrt{r^2 - (c/2)^2}$ ."
      },
      {
        "step": 2,
        "description": "Substitute the radius  $r = 10$  and the chord length  $c = 12$  into the formula."
      },
      {
        "step": 3,
        "description": "Calculate the distance from the center to the chord:  $d = \sqrt{10^2 - (12/2)^2} = \sqrt{100 - 36} = \sqrt{64} = 8$ ."
      }
    ],
    "final_answer": {
      "description": "The distance from the center to the chord is 8 units."
    }
  }
}
```

Dataset Exploration-[4]

(Text Data Distribution)



Dataset Exploration-[5]

(Image 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

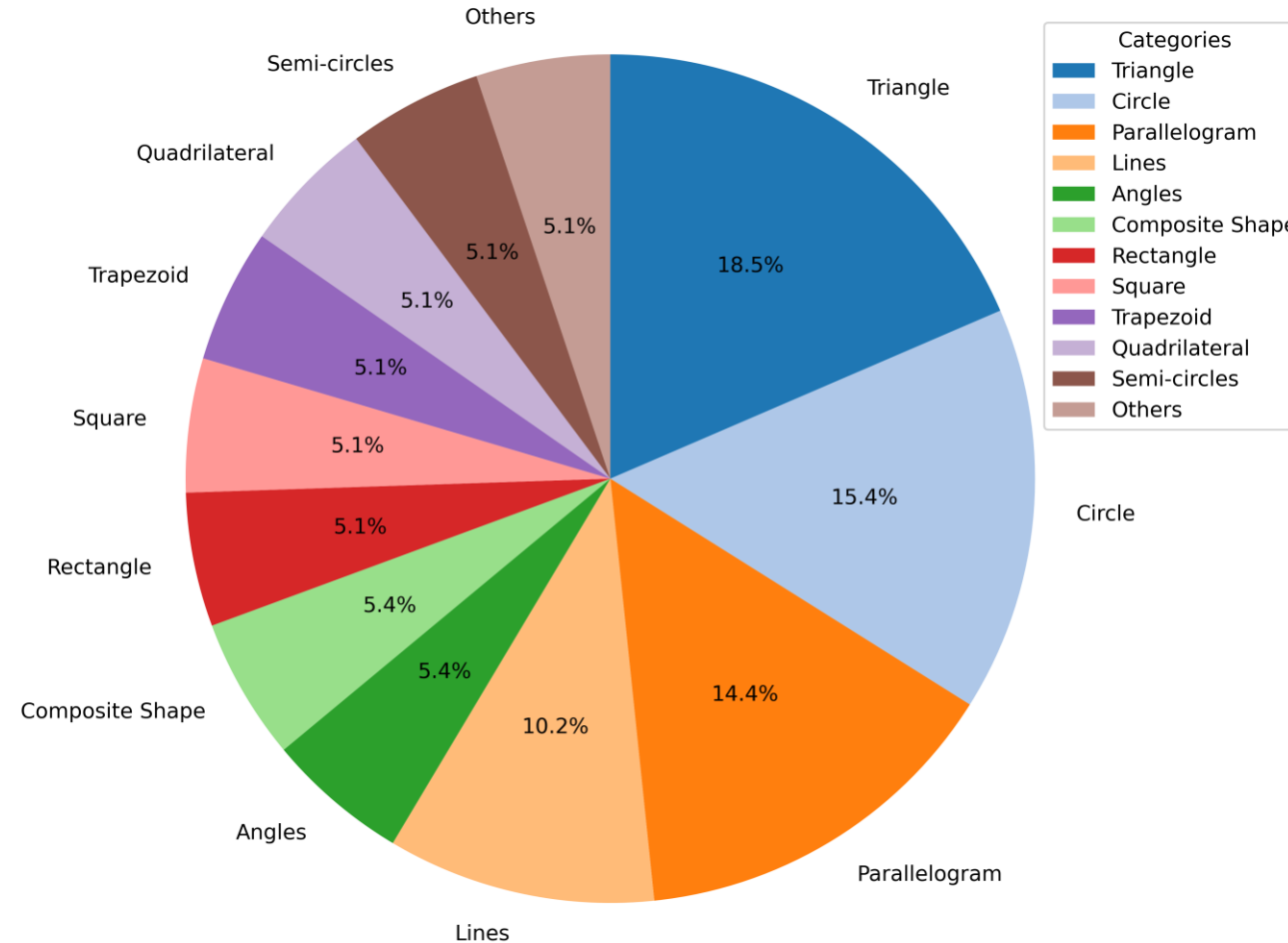
Dataset Exploration-[6] (Image Dataset Snippet)

8/9/2024

```
{
  "id": "20",
  "question": {
    "text": "Find the length of arc AB.",
    "image_url": "Image Dataset/batch_3/2.png"
  },
  "choices": [
    "10.85",
    "9.77",
    "12.35",
    "5.93"
  ],
  "answer": "9.77",
  "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:  $L = re$ ."
      },
      {
        "step": 3,
        "description": "Convert the central angle from degrees to radians:  $6 = 80^\circ * (\pi/180) = 1.39629$ ."
      },
      {
        "step": 4,
        "description": "Calculate the arc length:  $L \approx 1.39629 * 9.77403$ ."
      }
    ],
    "final_answer": {
      "description": "The length of arc AB is approximately 9.77 meters."
    }
  }
}
```

Dataset Exploration-[7]

(Image Data Distribution)



Result-[1] (Text Input)

Q. In a quadrilateral, two angles are 90 degrees and 60 degrees, and the third angle is 80 degrees. Find the measure of the fourth angle.

Actual Output

Steps:

Step 1: The sum of angles in a quadrilateral is 360 degrees.

Step 2: Given angles are 90, 60, and 80 degrees.

Step 3: Let x be the unknown angle.

Step 4: Therefore, $90 + 60 + 80 + x = 360$.

Step 5: Simplify to find x: $230 + x = 360$.

Step 6: Subtract 230 from 360 to get x: $x = 130$.

Final Answer: The measure of the fourth angle is 130 degrees.

Model's Output

The sum of angles in a quadrilateral is 360 degrees.

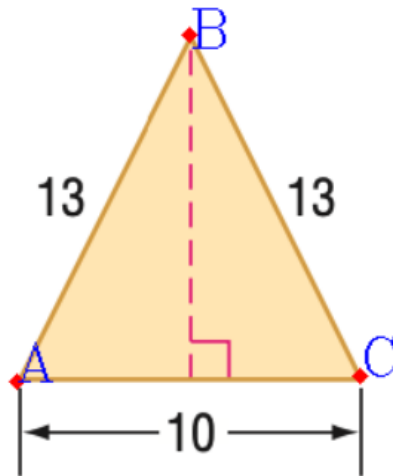
Given angles are 90, 60 and 80 degrees.

Let x be fourth angle.

On solving $x = 110$.

Result-[2] (Image Input)

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: $\text{Area} = 0.5 * \text{base} * \text{height}$. Substitute the known values: $\text{Area} = 0.5 * 10 * 12$.

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 base of 10 units and a height of 13 units. The area of the triangle can be calculated using the formula:

$$\text{Area} = (\text{Base} * \text{Height}) / 2$$

$$\text{Area} = (10 * 13) / 2$$

$$\text{Area} = 130 / 2$$

$$\text{Area} = 65$$

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

Discussion and Analysis- [1]

Benchmark

	MathVista (testmini)			GeoEval				
Model	GPS	GEO	AVG	GeoEval-2000 (A %)	GeoEval-2000 (T %)	GeoEval-backward (A %)	GeoEval-aug (A %)	GeoEval-hard (A %)
Human Performance	48.4	51.4	49.9	-	-	-	-	-
GPT-4V (Playground)	50.5	51.0	50.75	37.22	43.86	26.00	45.75	10.1
LLaVA-1.5-7B-hf	20.67	20.92	20.795	12.8	21.01	11.33	20.25	20.3

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

- **GeoViz Dataset Expansion and Fine-tuning**
 - GeoViz dataset development ongoing, with plans to expand to cover a wider range of mathematical topics and problem types.
 - Validation process refined and quality improved through iterative testing and feedback incorporation.
- **Fine-tuning LLaVa on GeoViz Dataset**
 - LLaVa fine-tuned on the expanded dataset to better understand and solve geometric problems.
- **Evaluation and Comparison of Fine-tuned LLaVa**
 - Model evaluated against other models or benchmarks to determine fine-tuning effectiveness and model performance.

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