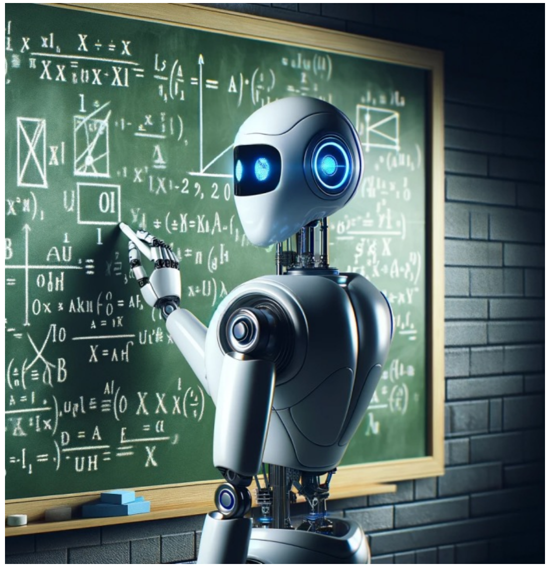
AI-Powered Mathematical Olympiad Problem Solver

Progress, Methods, and Future Directions

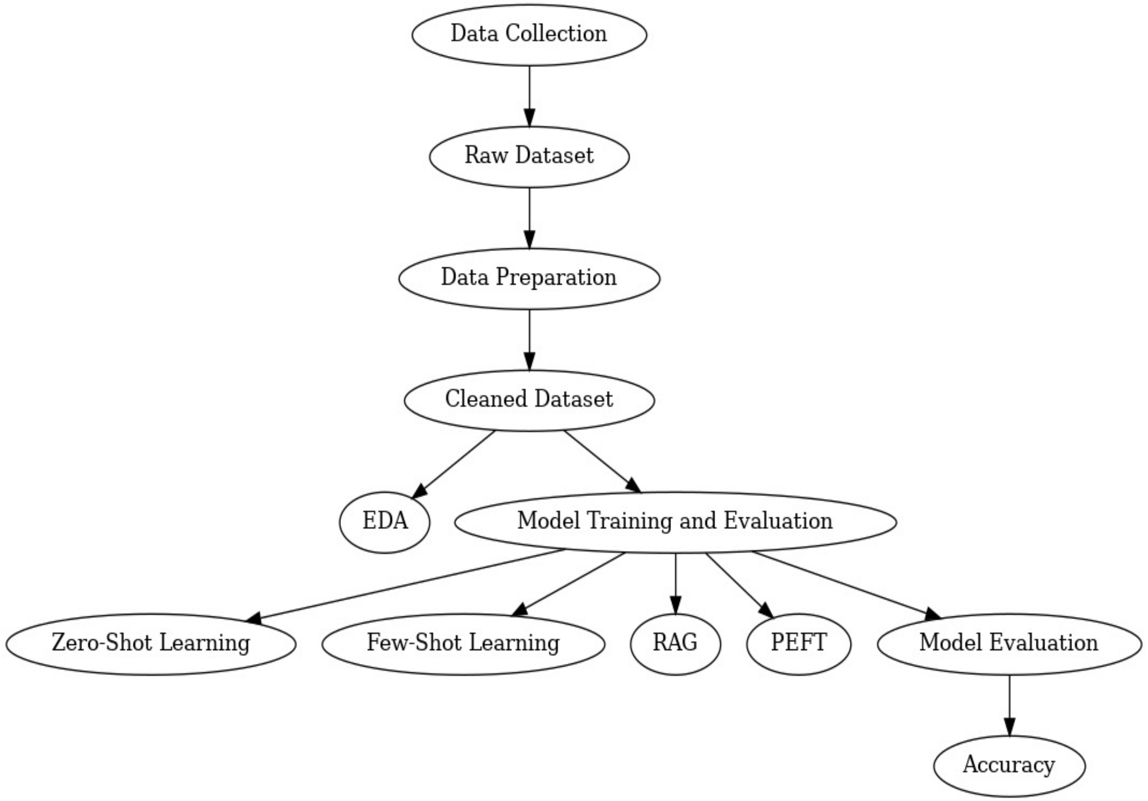
Introduction

- **Objective**:
  - **Goal**: Develop a model that accurately solves mathematical problems from the AI Mathematical Olympiad competition.
  - **Benchmark**: Exceed the performance of the current Gemma 7B benchmark of 3/50 and achieve at least 20% accuracy.
- **Key Components**:
  - **Problem Interpretation** : Develop a system that accurately interprets and understands natural language descriptions of Math Olympiad problems interpretation process.
  - **Mathematical Reasoning** : Implement algorithms to process and reason through mathematical concepts, theorems, and logic.
  - **Evaluation**: Establish a robust evaluation mechanism to assess the accuracy and reliability of the solutions provided by the model.
- **Expected Outcome**:By the end of this project, we anticipate having a robust model capable of solving a wide range of Math Olympiad problems with high accuracy. This model will not only serve as a powerful tool for students and educators but also pave the way for further advancements in AI-driven mathematical problem-solving.



Methodology

- **Data Collection:**
  - Gather: Comprehensive dataset of Math Olympiad problems and solutions.
  - Sources: Web scraping from AIME and Kaggle.
- **Data Preparation:**
  - Clean: Handle missing values, remove duplicates, and standardize formats.
  - Preprocess: Extract numerical answers and ensure data consistency.
- **Exploratory Data Analysis (EDA):**
  - Analyze: Perform summary statistics, distribution plots, and word clouds.
  - Visualize: Use histograms, word clouds, and n-gram frequencies.
- **Model Training and Evaluation:**
  - Techniques: Zero-Shot, Few-Shot, RAG, and PEFT.
  - Evaluate: Use accuracy



Dataset Collection

- **Sources**
  - The dataset has been compiled from reputable sources to ensure a diverse and comprehensive collection of Math Olympiad problems:
  - **Kaggle**: The dataset includes problems and solutions contributed by the Kaggle community, ensuring a wide range of difficulty levels and problem types.
  - **American Invitational Mathematics Examination (AIME):** Problems from AIME provide a high standard of difficulty and are essential for training the model to handle advanced mathematical concepts and techniques.
- **Focus**:
  - Problems and solutions formatted in LaTeX.

Data Preprocessing

- **Data merge**: Combined datasets into a single comprehensive dataset.
- **Normalization**: Standardized formatting and notation across all records to prevent inconsistencies that might confuse the model.
- **Question Filtering:** Selected questions with integer answers, discarding those without.
- **Answer Extraction:** Retrieved integer answers from the solutions for questions where answers were missing or not provided.
- **Cleaning:** Removed extraneous information, such as typographical errors and irrelevant content, to maintain the data's quality and relevance.

Dataset Structure

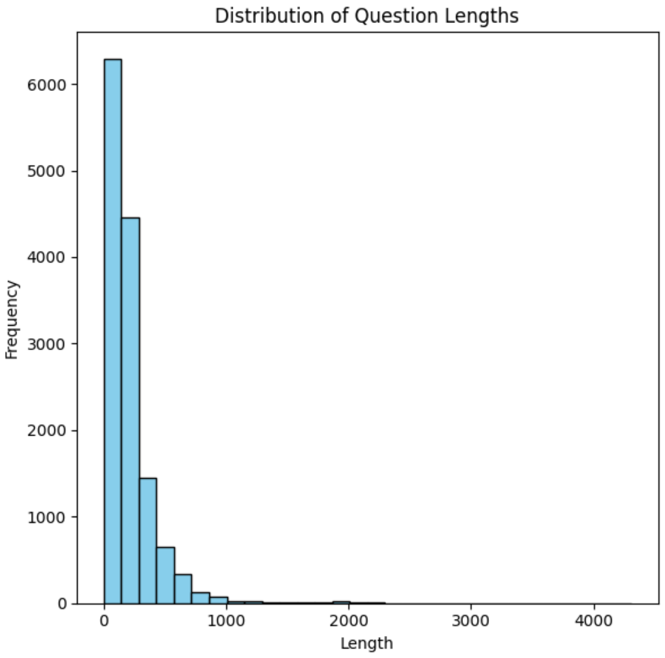
- The training dataset consists of approximately 9000 records, each containing three main components:
- **Questions:** These are the Math Olympiad problems presented in natural language. The questions cover various mathematical topics, including algebra, geometry, number theory, and combinatorics. Each question is designed to challenge the problem-solving abilities and conceptual understanding of the solver.
- **Solutions:** This column contains detailed solutions to the corresponding questions. The solutions are written in a step-by-step manner, illustrating the logical and mathematical reasoning required to arrive at the correct answer. These solutions are critical for training the LLM to understand the process of solving complex mathematical problems.
- **Answer**: The final integer answer

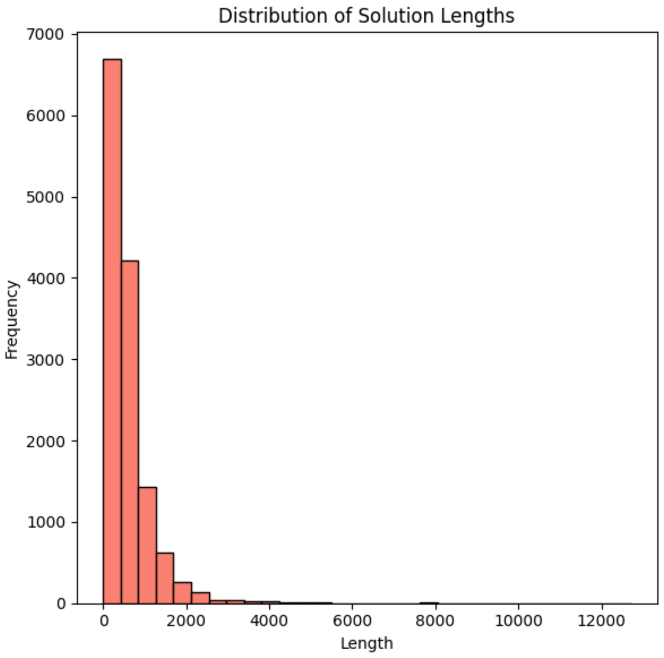
Exploratory Data Analysis (EDA)

- **Basic Information**: Total entries, unique questions, solutions, and answers.
- **Statistics**: Mean, standard deviation, minimum, and maximum lengths of questions and solutions.
- **Visualizations**: Histograms and word clouds.

EDA: Contd.

Distribution of question lengths and solution lengths.

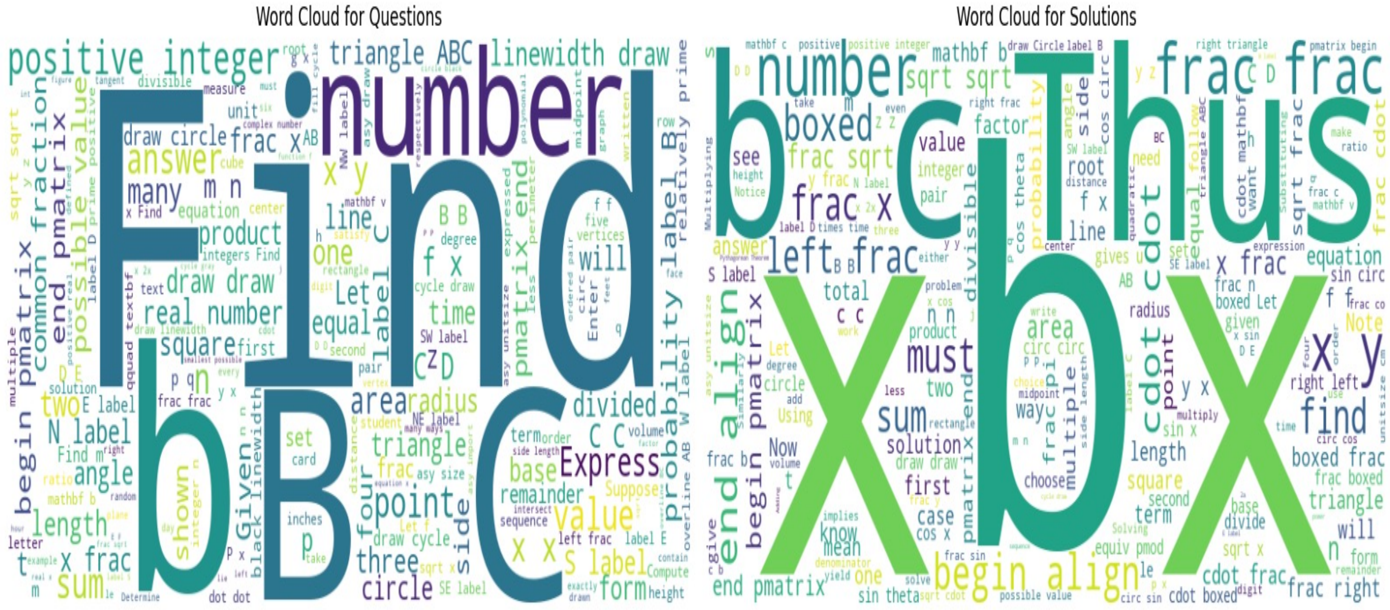




- **Question Length Distribution:**
  - Right-Skewed: Most questions are short, with a few very long ones.
  - High Frequency: Majority of questions are under 1000 characters.
- **Solution Length Distribution:**
  - Right-Skewed: Most solutions are concise, with some very long ones.
  - High Frequency: Majority of solutions are under 2000 characters.

EDA: Contd.

Common terms in the questions and Common terms in the solutions.



- **Questions:**
  - Common Terms: “Find,” “number,” “integer,” “positive,” and “real” are the most frequent words.
  - Focus: Emphasis on finding numbers, integers, and specific values.
- **Solutions**:
  - Common Terms: “Thus,” “find,” “number,” “frac,” and “cdot” are the most frequent words.
  - Focus: Emphasis on explaining solutions with mathematical terms and operations.

Methods and Accuracy on 5 Complex Olympiad Problems

- **Zero-Shot Learning:**
  - **Description**: Uses a pre-trained model without additional training on specific examples.
  - **Accuracy**: 0%
  - **Note**: The steps were correct in most, but the execution were not leading to incorrect result.
- **Few-Shot Learning:**
  - **Description**: Uses a few examples from AIME to guide predictions.
  - **Accuracy**: 0%
  - No improvement, indicating the need for more relevant examples or techniques.
- **Retrieval-Augmented Generation (RAG):**
  - **Description**: Uses BERT and FAISS to retrieve relevant examples and augment input.
  - **Accuracy**: 10%
  - Showed improvement by leveraging retrieved information, resulting in one correct answer and a better reasoning.
- **Reason for Using 5 Data Points:** Using 5 data points allows for quicker iterations (~1min/problem) and faster identification of issues due to time constraints. It is practical to start with a manageable number of complex problems, helping to set a baseline before scaling to larger datasets.

Directions Checkpoint II

- **Fine-Tuning with PEFT**: Plan to use transfer learning and fine-tune the deepseekmath model for better performance.
- **Experiment Other SOA Methods**:.
  - **Augmentation:** Generating additional training examples through data augmentation techniques, such as paraphrasing questions and rephrasing solutions, to enhance the model's generalization capabilities.
  - **CoT and ToT**: Use logical reasoning chains and tree structures to break down complex problems.
  - **Ensemble Methods**: Combining multiple models to improve accuracy.

Reference

- AIMO challenge: https://www.kaggle.com/competitions/ai-mathematical-olympiad-prize
- AIME: https://artofproblemsolving.com/wiki/index.php/2018_AIME_I_Problems/Problem_3
- https://dassum.medium.com/fine-tune-large-language-model-llm-on-a-custom-dataset-with-qlora-fb60abdeba07
- https://github.com/ZeusSama0001/RAG-chatbot/blob/main/model.py
- https://www.kaggle.com/code/awsaf49/aimo-kerasnlp-starter