Technical Appendix

Reproducibility Checklist

This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced: **yes**
- Clearly delineates statements that are opinions, hypotheses, and speculation from objective facts and results: **yes**
- Provides well-marked pedagogical references for less-familiar readers to gain the background necessary to replicate the paper: **yes**

Does this paper make theoretical contributions? No
Does this paper rely on one or more datasets? yes

- A motivation is given for why the experiments are conducted on the selected datasets: **yes**
- All novel datasets introduced in this paper are included in a data appendix: **partial**
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes: **yes**
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations: **yes**
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available: **yes**
- All datasets that are not publicly available are described in detail, with an explanation of why publicly available alternatives are not scientifically satisfactory: **partial**

Does this paper include computational experiments? yes

- Any code required for pre-processing data is included in the appendix: **partial**
- All source code required for conducting and analyzing the experiments is included in a code appendix: **partial**
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes: **yes**
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from: **partial**
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results: **no**
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks: **no**
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics: **no**
- This paper states the number of algorithm runs used to compute each reported result: **yes**
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information: **no**
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank): **no**
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments: **partial**
- This paper states the number and range of values tried per (hyper-)parameter during the development of the paper, along with the criterion used for selecting the final parameter setting: **no**

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Model Parameters and Data Explanation

Model Methodology Overview

Our research employed an advanced sequence inference technique to fine-tune our model for generating responses to specific prompts. These prompts were meticulously crafted with clear instructions and sample questions to ensure the model produced accurate and varied responses.

Hyper parameters Used:

- Sampling: Enabled
- Top_k: 40 (for relevance and diversity)

- Temperature: 0.8 (for balanced creativity)
- Top_p: 0.90 (for refined nucleus sampling)
- Maximum Length: 4096 (for comprehensive answers)
- Fine-tuning: 3 epochs

Fine-Tuning Prompt

Definition: This prompt is designed to fine-tune a model by providing a specific instruction and a corresponding response. The focus is on ensuring that the model accurately completes the task while being cautious of incorrect calculations and avoiding repetition of errors.

```
Below is an instruction that describes a
task. Write a response that appropriately
completes the request. Be aware of wrong
calculations and do not repeat them.

### Instruction:
{sample['Question']}

### Response:
{sample['Response']}
```

Data Augmentation Prompt

Definition: This prompt is used for generating new, conceptually similar questions and answers based on an existing problem. It involves creating problems with varied difficulty levels to enhance the diversity and complexity of the dataset.

```
Your task is to create a similar concept-
-ual question and answer with diverse
difficulty levels (either similarly simp-
-le, the same, or more complex) using
the provided problem.

Problem:
Question: {sample['Example']}
Answer: {sample['refined_solution']}

New Problem: {sample['Question']}
```

Decomposition Strategy Prompt

Definition: This prompt is used for generating updated solutions of HAWP dataset using the Decomposition Strategy, for questions involving multiplication and division. HAWP contains single-operator word problems, so we divided each problem by the type of operation used in the solution. We used separate prompts for Multiplication problems and Division problems. We also gave examples to support the generation.

```
Your task is modify the following mathem-
-atical solution by breaking down the mu-
-ltiplicand into place value components
(hundreds, tens, ones, etc.) and then
multiplying each component by the other
multiplicand. Then, sum the products
to get the final result.

Answer: {sample['solution']}

New Answer: {sample['updated_solution']}
```

One of the examples provided as few-shot example is given below:

```
Answer: 543 × 27 = 14661

Updated Answer:
Break down 543 into place value components:

543 = 500 + 40 + 3

Multiply each component by 27:

500 × 27 = 13500

40 × 27 = 1080

3 × 27 = 81

Add the products:
13500 + 1080 + 81 = 14661

Final Answer: 543 multiplied by 27
equals 14661.
```

The Decomposition Strategy prompt for division is given below

```
Your task is modify the following mathem-
-atical solution by decomposing the
dividend into segments, then divide
each by the divisor, and sum the
quotients to obtain the final answer.

Answer: {sample['solution']}

New Answer: {sample['updated_solution']}
```

One of the examples provided as few-shot example is given below:

```
Answer: 968 ÷ 16 = 60.5

Updated Answer:
Break down 968 into place value
components:

968 = 800 + 160 + 8

Divide each component by 16:

800 ÷ 16 = 50

160 ÷ 16 = 10

8 ÷ 16 = 0.5
Add the quotients:
50 + 10 + 0.5 = 60.5

Final Answer: 968 divided by 16
equals 60.5.
```

Explanation of Key Terms and Methodology Components

In our study, we developed a structured approach to provide solutions for each raw question-answer pair, consisting of the following steps:

Data Identification: Specifying the relevant data needed for the problem. This includes identifying variables, constants, and any conditions or parameters related to the problem.

Problem Analysis: Examining the problem to understand its components and determine suitable methods for solving it. This includes understanding theoretical aspects such as set theory or integration rules.

Theoretical Framework: Establishing the foundational theories and principles that guide the solution. Examples include using set operations for probability problems or antiderivative rules for integrals.

Methodology Development: Creating a detailed, step-by-step plan to solve the problem. This includes developing necessary procedures and formulae for the solution.

Computation: Performing the calculations and applying the developed methodology to derive the final solution.

Solution: Presenting the final answer, ensuring it is clear and easy to understand.

Data Classification: Easy, Medium, and Hard

To classify data into easy, medium, and hard categories, we use the following criteria:

1. **Language Understanding:** Assesses the complexity of the language used. Problems with straightforward, clear language are classified as easy, while those with technical or complex phrasing are considered medium or hard.
2. **Mathematical Complexity:** Assesses the nature of mathematical operations required. Problems involving fundamental concepts are easy, whereas those requiring multiple or intricate operations are medium or hard.
3. **Reasoning Complexity:** Evaluates the complexity of reasoning needed. Problems requiring straightforward application of concepts are easy, while those needing detailed logic or multi-step reasoning are medium or hard.
4. **Number of Variables:** Evaluates the number of variables involved. Problems with a few variables are easy, whereas those with many or complex variables are medium or hard.
5. **Conceptual Complexity:** Assesses the depth of understanding required. Basic conceptual understanding is easy, while advanced or abstract concepts are medium or hard.

Structured Problem Solving: Examples of Data Transformation

We have raw data in the format of **Question** and **Final Answer**, which is then converted into a **Refined Solution**. These examples span various topics, including but not limited to Probability, Linear Algebra, Integration, and many more. Following are examples of few instances from the curated IndiMathQA dataset.

Example 1: Conic Sections

Question: Find the equation of the hyperbola with foci $(0, \pm 3)$ and vertices $\left(0, \pm \frac{\sqrt{11}}{2}\right)$.

Raw Solution: Given the foci on the y-axis, the equation of the hyperbola is of the form

$$\frac{y^2}{a^2} - \frac{x^2}{b^2} = 1.$$

Using the given vertices $\left(0, \pm \frac{\sqrt{11}}{2}\right)$, we deduce that

$$a = \frac{\sqrt{11}}{2}.$$

The foci are $(0, \pm 3)$, which means

$$c = 3.$$

Using the relationship $b^2 = c^2 - a^2$, we find

$$b^2 = 3^2 - \left(\frac{\sqrt{11}}{2}\right)^2 = 9 - \frac{11}{4} = \frac{36}{4} - \frac{11}{4} = \frac{25}{4}.$$

This yields the equation of the hyperbola as

$$\frac{y^2}{\left(\frac{11}{4}\right)} - \frac{x^2}{\left(\frac{25}{4}\right)} = 1,$$

which simplifies to

$$100y^2 - 44x^2 = 275.$$

Refined Solution:

Data Identification: - Foci: $(0, \pm 3)$ - Vertices: $\left(0, \pm \frac{\sqrt{11}}{2}\right)$

Problem Analysis: We need to determine the standard form of the equation of the hyperbola given its foci and vertices. For a hyperbola centered at the origin and oriented along the y-axis, its general equation is

$$\frac{y^2}{a^2} - \frac{x^2}{b^2} = 1,$$

where: - The distance from the center to a vertex is a. - The distance from the center to a focus is c. - The relationship $c^2 = a^2 + b^2$ holds true.

Theoretical Framework: Key concepts and formulae for hyperbolas include: - Standard form of hyperbola equation: $\frac{y^2}{a^2} - \frac{x^2}{b^2} = 1$ - Distance to the vertices: $\pm a$ - Distance to the foci: $\pm c$ - Relationship: $c^2 = a^2 + b^2$

Methodology Development: From the given foci and vertices, we can identify a and c. Using the relationship $c^2 = a^2 + b^2$, we can find b and formulate the hyperbola's equation.

Computation:

1. Identify c:

$$c = 3$$

2. Identify a:

$$a = \frac{\sqrt{11}}{2}$$

3. Apply the relationship $c^2 = a^2 + b^2$:

$$c^2 = a^2 + b^2 \implies 3^2 = \left(\frac{\sqrt{11}}{2}\right)^2 + b^2$$

4. Compute a^2:

$$a^2 = \left(\frac{\sqrt{11}}{2}\right)^2 = \frac{11}{4}$$

5. Solve for b^2:

$$9 = \frac{11}{4} + b^2 \implies 9 - \frac{11}{4} = b^2$$

$$\implies \frac{36}{4} - \frac{11}{4} = b^2 \implies \frac{25}{4} = b^2$$

Therefore,

$$b^2 = \frac{25}{4}$$

Therefore,

$$b^2 = \frac{25}{4}$$

Solution: Using the values of a^2 and b^2, the equation of the hyperbola is: Using the values of a^2 and b^2, the equation of the hyperbola is:

$$\frac{y^2}{\left(\frac{\sqrt{11}}{2}\right)^2} - \frac{x^2}{\left(\frac{5}{2}\right)^2} = 1$$

$$\implies \frac{y^2}{\frac{11}{4}} - \frac{x^2}{\frac{25}{4}} = 1$$

$$\implies \frac{4y^2}{11} - \frac{4x^2}{25} = 1$$

Simplifying by multiplying through by 100 to clear the denominators:

$$100y^2 - 44x^2 = 275$$

Hence, the equation of the hyperbola is:

$$\boxed{100y^2 - 44x^2 = 275}$$

Example 2: Probability

Question: A committee of two persons is selected from two men and two women. What is the probability that the committee will have (a) no man? (b) one man? (c) two men?

Raw Solution: The total number of persons $= 2 + 2 = 4$. Out of these four persons, two can be selected in ${}^4\mathrm{C}_2$ ways.

(a) No men in the committee of two means there will be two women in the committee.

Out of two women, two can be selected in ${}^2\mathrm{C}_2 = 1$ way.

Therefore,

$$\mathrm{P}(\text{no man}) = \frac{{}^2\mathrm{C}_2}{{}^4\mathrm{C}_2} = \frac{1 \times 2 \times 1}{4 \times 3} = \frac{1}{6}$$

(b) One man in the committee means that there is one woman. One man out of 2 can be selected in ${}^2\mathrm{C}_1$ ways and one woman out of 2 can be selected in ${}^2\mathrm{C}_1$ ways. Together they can be selected in ${}^2\mathrm{C}_1 \times {}^2\mathrm{C}_1$ ways.

Therefore,

$$\mathrm{P}(\text{One man}) = \frac{{}^2\mathrm{C}_1 \times {}^2\mathrm{C}_1}{{}^4\mathrm{C}_2} = \frac{2 \times 2}{2 \times 3} = \frac{2}{3}$$

(c) Two men can be selected in ${}^2\mathrm{C}_2$ way.

Hence,

$$\mathrm{P}(\text{Two men}) = \frac{{}^2\mathrm{C}_2}{{}^4\mathrm{C}_2} = \frac{1}{{}^4\mathrm{C}_2} = \frac{1}{6}$$

Final Answer:

- The probability that the committee will have no man is $\frac{1}{6}$.
- The probability that the committee will have one man is $\frac{2}{3}$.
- The probability that the committee will have two men is $\frac{1}{6}$.

Refined Solution: Data Identification:

- Total people: 2 men and 2 women.
- Committee size: 2 persons.

Problem Analysis:

- Calculate the total number of ways to form a committee of 2 persons from 4 people.
- Determine the number of favorable outcomes for:
  - No man in the committee.
  - One man in the committee.
  - Two men in the committee.
- Calculate the probability for each case by dividing the number of favourable outcomes by the total number of ways to form a committee.

Theoretical Framework:

- Use the combination formula $C(n, r)$ which denotes the number of ways to choose r items from n items without regard to the order, given by $C(n, r) = \frac{n!}{r!(n-r)!}$.
- Probability of an event $P(E) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}$.

Methodology Development:

1. Total number of ways to form a committee of 2 out of 4 people: $C(4, 2)$.
2. Number of ways to form a committee with:
   - No man: $C(2, 2)$ (choosing 2 women out of 2).
   - One man: $C(2, 1) \times C(2, 1)$ (choosing 1 man out of 2 and 1 woman out of 2).
   - Two men: $C(2, 2)$ (choosing 2 men out of 2).

Computation:

1. Total number of ways to form the committee:

$$C(4, 2) = \frac{4!}{2!(4-2)!} = \frac{4 \times 3}{2 \times 1} = 6$$

2. Number of favorable outcomes:
   - No man:
$$C(2,2) = \frac{2!}{2! \times 0!} = 1$$
   - One man:
$$C(2,1) \times C(2,1) = \left(\frac{2!}{1! \times (2-1)!}\right) \times \left(\frac{2!}{1! \times (2-1)!}\right) = 2 \times 2 = 4$$
   - Two men:
$$C(2,2) = \frac{2!}{2! \times 0!} = 1$$
3. Calculate the probabilities:
   - No man:
$$P(\text{No man}) = \frac{\text{Number of outcomes for no man}}{\text{Total number of ways}} = \frac{1}{6}$$
   - One man:
$$P(\text{One man}) = \frac{\text{Number of outcomes for one man}}{\text{Total number of ways}} = \frac{4}{6} = \frac{2}{3}$$
   - Two men:
$$P(\text{Two men}) = \frac{\text{Number of outcomes for two men}}{\text{Total number of ways}} = \frac{1}{6}$$

Solution:

- The probability that the committee will have no man is $\frac{1}{6}$.
- The probability that the committee will have one man is $\frac{2}{3}$.
- The probability that the committee will have two men is $\frac{1}{6}$.

These refined solutions provide clear, step-by-step explanations of the problem, ensuring the process is thoroughly documented and easier to follow. This structured approach enhances the accuracy and clarity of solution generation using large language models (LLMs).

Consistency of Topics across Difficulties

In this section, we demonstrate that during our dataset preparation and categorization of questions into Easy, Medium, and Hard levels, we ensured consistent representation of mathematical topics across all complexity levels in both the MATHS and IndiMath datasets. Figure 1 and 2 illustrates the uniform distribution of each mathematical topic at each difficulty level, highlighting the LLM's ability to effectively learn and generalize across topics from easy to hard levels. remove Math dataset

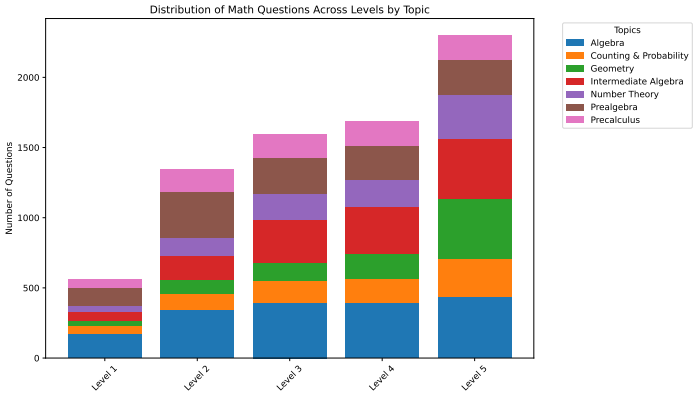

Figure 1: Bar graph showing topic distribution across each difficulty level extracted from the MATH dataset

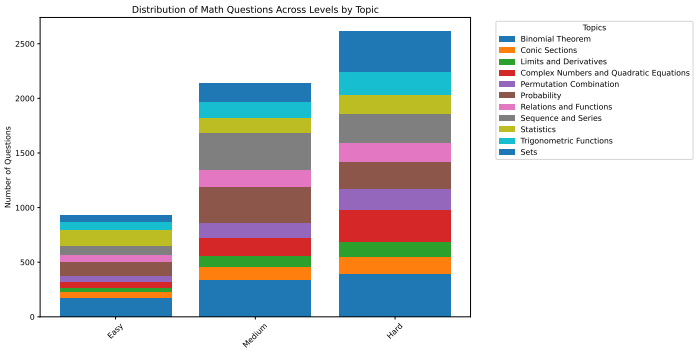

Figure 2: Bar graph showing topic distribution across each difficulty level extracted from the IndiMathQA dataset

Curriculum Learning Analysis

In this section, we present heatmap visualizations showing each model's final (best) performance for easy, medium, and hard questions in both languages, achieved through our experimentations. Figure 3 and 4 show that WizardMath consistently outperforms other open-source counterparts across all difficulties, while LLeMMA performs the worst. These results were obtained after we were able to counter the hallucination problem in LLeMMA to some extent through our strategies as discussed in the Ablation Study (Section 4) of the Main Paper.

Error Analysis: Demonstrating the effect of Curriculum Learning

Here we present the answers of the WizardMath 7B model before and after SFT based on Curriculum Learning.

- **Question**: A family has two children. What is the probability that both children are boys given that at least one of them is a boy?
- **Wizard Math Base Model Response**: To solve the problem of finding the probability that both children are boys given that at least one of them is a boy, we assume that if one child is a boy, the other must also

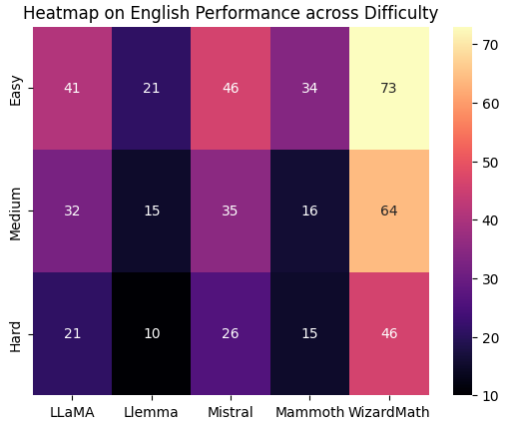

Figure 3: Heatmap showing English mathematical performance across Difficulties (Easy, Medium, Hard) on each Mathematical Models.

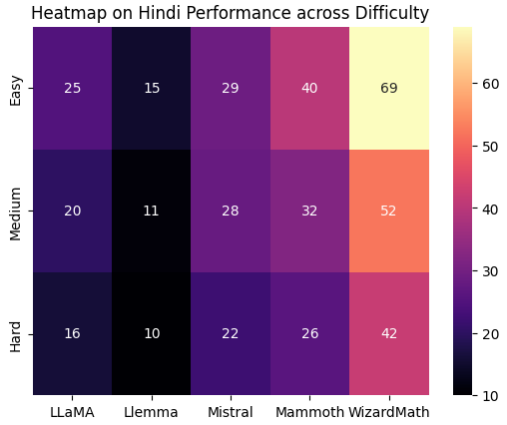

Figure 4: Heatmap showing Hindi mathematical performance across Difficulties (Easy, Medium, Hard) on each Mathematical Model.

be a boy. Therefore, the probability that both children are boys is 100%.

- **Wizard Math SFT_Easy Response**: To solve the problem of finding the probability that both children are boys given that at least one of them is a boy, we use conditional probability. There are four possible combinations for two children: (Boy, Boy), (Boy, Girl), (Girl, Boy), and (Girl, Girl). Since we know at least one child is a boy, we eliminate the (Girl, Girl) scenario, leaving us with (Boy, Boy), (Boy, Girl), and (Girl, Boy). Each of these scenarios has an equal probability of 1/4. Thus, the probability of at least one child being a boy is 3/4. The probability that both children are boys given this condition is calculated as the probability of both being boys (1/4) divided by the probability of at least one being a boy (3/4), resulting in 1/2. Therefore, the probability that both children are boys given that at least one is a boy is 1/2.
- **Wizard Math SFT_Easy+Medium Response**: To solve the problem of finding the probability that both children are boys given that at least one of them is a boy, we use conditional probability. There are four possible combinations for two children: (Boy, Boy), (Boy, Girl), (Girl, Boy), and (Girl, Girl). Since we know at least one child is a boy, we eliminate the (Girl, Girl) scenario, leaving us with (Boy, Boy), (Boy, Girl), and (Girl, Boy). Each of these scenarios has an equal probability of 1/4. Thus, the probability of at least one child being a boy is 3/4. The probability that both children are boys given this condition is calculated as the probability of both being boys (1/4) divided by the probability of at least one being a boy (3/4), resulting in 1/3. Therefore, the probability that both children are boys given that at least one is a boy is 1/3.
- **Comments**: The base model answer demonstrates its inability to apply probability logic correctly. After Curriculum learning with SFT on only Easy examples, the model's reasoning improved substantially, but the model faced an error in calculation at the very last step. This error was rectified when the model was fine-tuned on Easy+Medium, demonstrating the benefit of Curriculum Learning in teaching the LLM the basics of Math, like logical reasoning and calculations, before solving complex questions.

Annotation Analysis IndiMathQA

Our augmented dataset comprises question-answer pairs generated by GPT-4 and verified by five human annotators. This section provides an analysis of the annotation process and the consistency among annotators.

Evaluation Methodology and Inter-Annotator Agreement for GPT-4 Generated Responses

The evaluation methodology for GPT-4 generated responses involved a rigorous review process conducted by five independent human annotators. Each question-answer pair was meticulously evaluated for accuracy, relevance, and clarity. The verification process encompassed all aspects of the data, including data identification, problem analysis, theoretical framework, methodology development, computation, and the final solution. To ensure the reliability of these annotations, inter-annotator agreement was assessed using Fleiss' Kappa, focusing on two critical aspects.

The first aspect was the validation of solution correctness, where the annotators assessed whether the generated solution appropriately matched the given question. The Fleiss' Kappa score for this validation was 0.71, indicating substantial agreement among the annotators and confirming the reliability of the model's output. The second aspect involved classifying the data into easy, medium, and hard categories based on criteria such as language understanding, mathematical complexity, reasoning complexity, number of variables, and conceptual complexity. The Fleiss' Kappa score for this classification task was 0.58, reflecting moderate agreement. This score suggests that while there was general consensus, some variation existed among the annotators, likely due to subjective interpretations of what constitutes a medium

versus a hard problem. Such differences may arise from personal preferences or individual experiences with similar tasks. These Kappa scores collectively validate the robustness of the annotation process, demonstrating a high level of consistency in the human evaluation used to verify and classify GPT-4's generated responses.

Discrepancies and Resolutions

Discrepancies among the annotators' evaluations were resolved through a majority voting system. If a majority decision could not be reached due to close scores or ambiguity, a domain expert reviewed the responses to make the final decision. If the domain expert could not reach a consensus, the solution was manually corrected and re-evaluated.

Ethics Statement

The datasets used for training and testing the LLM were sourced from publicly available repositories. We acknowledge the potential for bias inherent in language model training datasets. The collected data comes from NCERT textbooks, that are used across India for high school education, hence representing the extended demographic within India.

Limitations

A limitation of our study is that we do not evaluate on all the available sources of data available. Future iterations of this research aim to expand the scope of evaluation, with the help of the research community to collect more comprehensive data from various available sources. Additionally, our study did not include the examination of romanized Hindi sentences, where Hindi word for elephant are written using the English alphabet as "Hathi." This form of input is prevalent in India, particularly when typing on electronic devices. Future research could beneficially extend to enhancing model performance on such inputs as well. The Easy/Medium/Hard distribution of questions was done manually, which can add bias to the distribution as we increase the dataset size.