



MIND: Math Informed syNthetic Dialogues for Pretraining LLMs

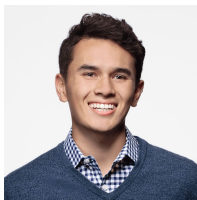
ICLR 2025



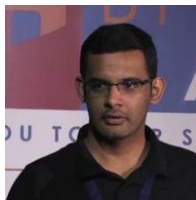
Syeda Nahida
Akter



Shrimai
Prabhumoye



John Kamalu



Sanjeev
Satheesh



Eric Nyberg



Mostofa
Patwary



Mohammad
Shoeybi



Bryan
Catanzaro

Why should we care about Math Reasoning?

Motivation

Conceptual Understanding

Math reasoning helps build a deeper understanding of mathematical concepts rather than just memorizing procedures

Problem-Solving Skills

Individuals can analyze problems, identify relevant information, and develop effective strategies to solve them.

Critical Thinking

Fosters critical thinking skills, allowing individuals to evaluate information, identify inconsistencies, and make informed decisions

Real World Applications

Equips individuals to apply math concepts to real-life situations



Why Synthetic Data?

Motivation

- Benchmarks that require logical and mathematical reasoning capabilities such as MMLU-Stem, MATH and GSM8K have been challenging for LLMs.
- Curating high-quality data is expensive.
- Distribution of high-quality data sources is very skewed.

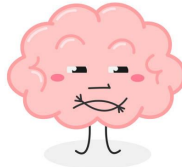


MIND: Math Informed syNthetic Dialogues

Proposed Approach

To upsample scarce high-quality data, we introduce a novel synthetic data generation approach:

Math Informed syNthetic Dialogues



MIND



Why MIND: Math Informed syNthetic Dialogues?


Motivation


1. Breaks down the original context step-by-step and generates multi-turn conversations


Raw Text


How many numbers are there between 20000 and 30000 in which the digits are 2,3,5,6,7 and each digit can be repeated any number of times. Since the numbers are > 20000 and < 30000 1st place can be only 2. Remaining 4 places can be occupied by any of the 5 digits. Therefore, Required no. of numbers = $5^4 = 625$


Conversation


 I don't understand the problem. Can you explain what it's asking?


 The problem is asking us to find the number of numbers between 20000 and 30000 that can be formed using the digits 2, 3, 5, 6, and 7. Each digit can be repeated any number of times.


 Okay, got it. So, why is the first place limited to only 2?

 That's because the numbers are greater than 20000 and less than 30000. *If the first digit were 1, the number would be less than 20000, and if it were 3, the number would be greater than 30000. So, the first place can only be 2.*

 I see. What about the remaining places? Can they be occupied by any of the given digits?

 Yes, that's correct. The remaining 4 places can be occupied by any of the 5 digits: 2, 3, 5, 6, and 7.

 Okay, so how do we find the total number of numbers that can be formed?

 Since each of the remaining 4 places can be occupied by any of the 5 digits, we have 5 options for each place. To find the total number of numbers, we raise 5 to the power of 4, which is $5^4 = 625$.

2. Leverages knowledge imbalances between participants for further expansion of the corpora



Student



Problem Solver



Teacher



Layman

Shared Knowledge Gap



Student



Problem Solver



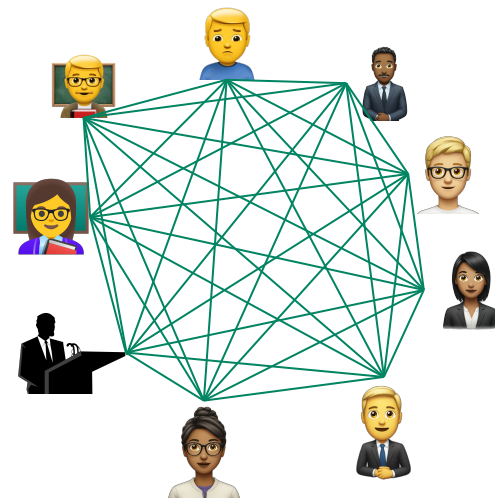
Student



Know-All

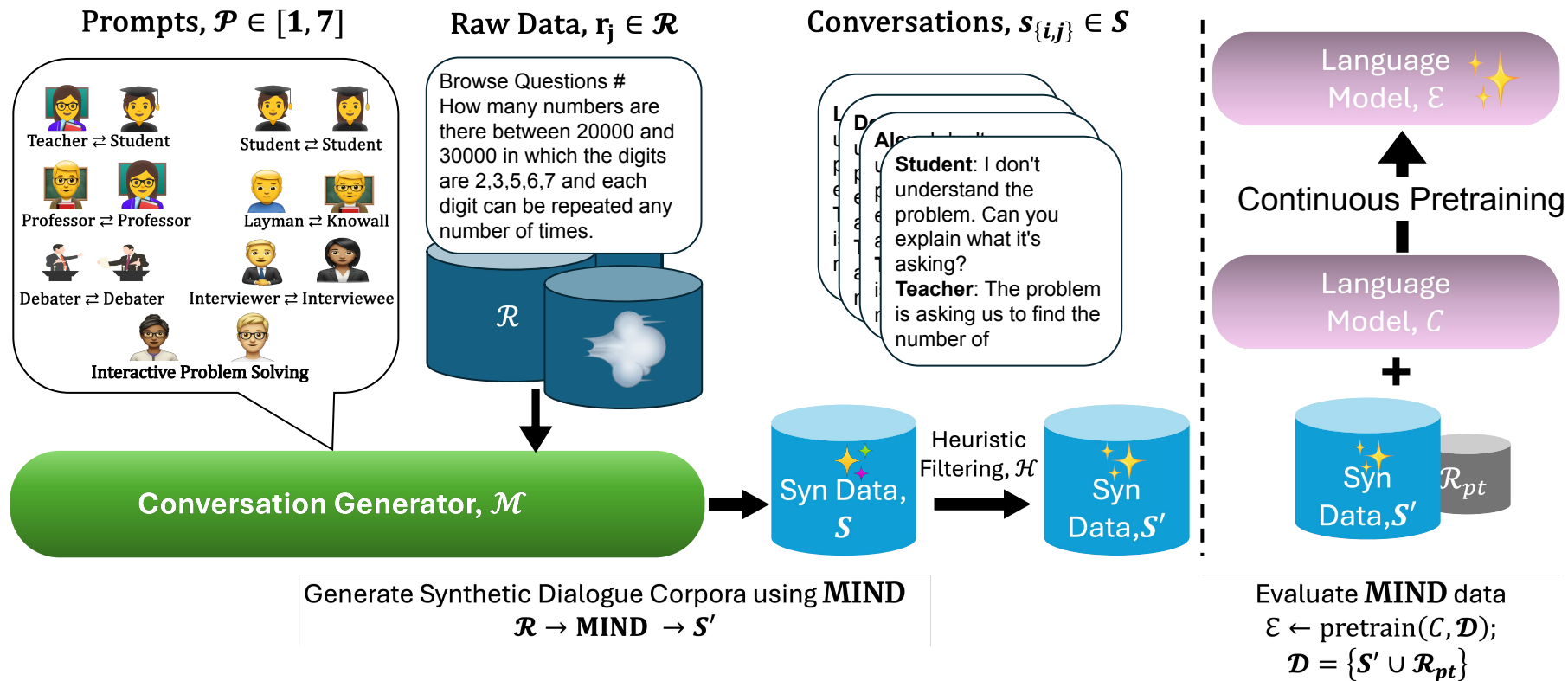
Knowledge Imbalance

3. Enables the continuous generation of synthetic data from a document by employing infinite conversational styles



How does MIND work?

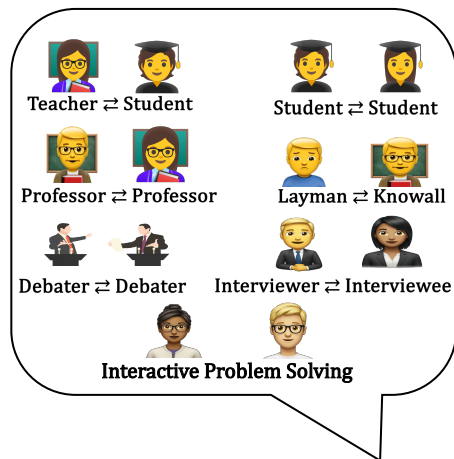
Proposed Framework



How does MIND work?

Proposed Framework

Prompts, $\mathcal{P} \in [1, 7]$



Conversation Generator, \mathcal{M}

Generate Synthetic Dialogue Corpora using **MIND**

$\mathcal{R} \rightarrow \text{MIND} \rightarrow \mathcal{S}'$



Dialogue Prompts

Methodology



Interviewer ⇌ Interviewee



Conduct an interview-style conversation where one participant acts as the interviewer, asking questions exclusively related to the content provided, while the other participant serves as the subject matter expert, providing **detailed responses** based on the content.



Debater ⇌ Debater



Convert the context above as a **multi-turn debate-style conversation** where the participants present **arguments and counterarguments** based solely on the content provided, without introducing external information or personal opinions. Each participant defends others arguments **step-by-step with chain-of-thoughts**.

Dialogue Prompts

Methodology



Knowall ⇌ Layman



Imagine you are presenting the content above step-by-step to a layman. While you are presenting, the layman has a lot of **follow up questions** regarding your presentation. You **answer the questions step-by-step with chain-of-thoughts**. Design this interaction between you and the layman as a multi-turn conversational manner.



Problem Solving

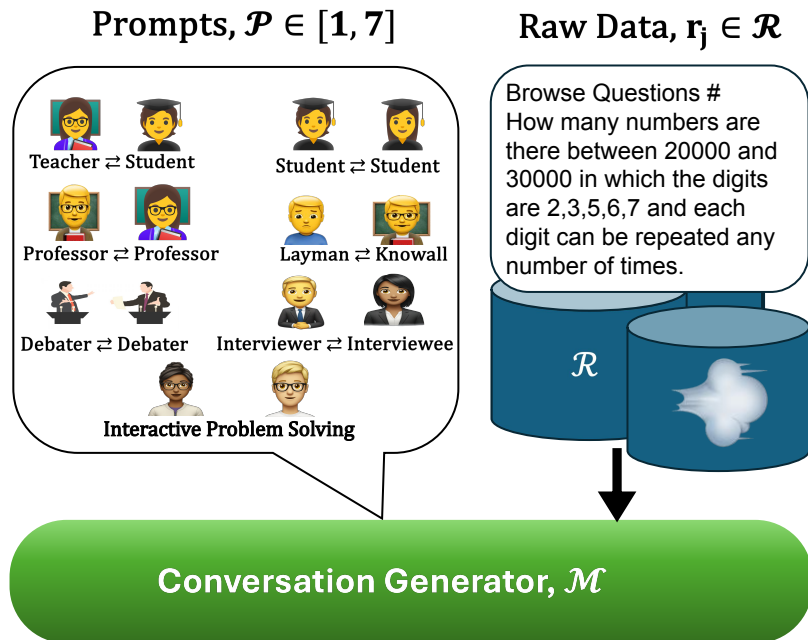


Convert the context above as a **multi-turn problem-solving** conversation where participants analyze challenges or scenarios presented in the content and brainstorm solutions within the context of the provided material, avoiding speculation or unrelated discussions.



How does MIND work?

Proposed Framework



Generate Synthetic Dialogue Corpora using MIND

$$\mathcal{R} \rightarrow \text{MIND} \rightarrow \mathcal{S}'$$

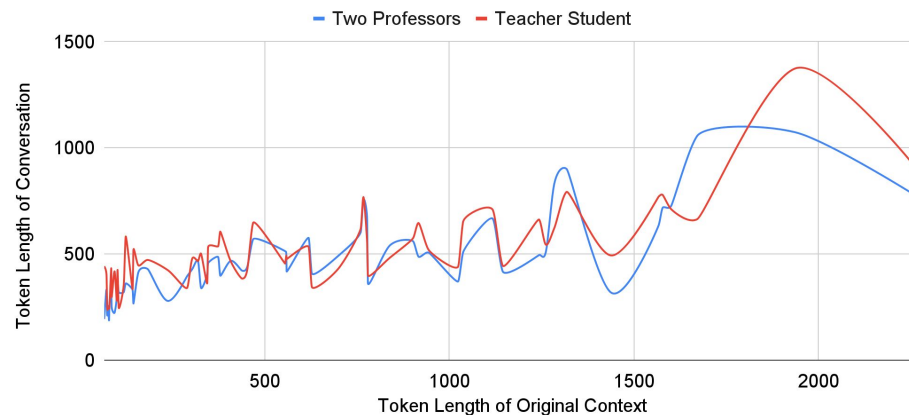


Data and LLM

Proposed Framework

$$s_{i,j} = \mathcal{M}(p_i || r_j) \text{ where } \mathcal{R} = \{r_1, \dots, r_N\} \text{ and } \mathcal{P} = \{p_1, \dots, p_7\}$$

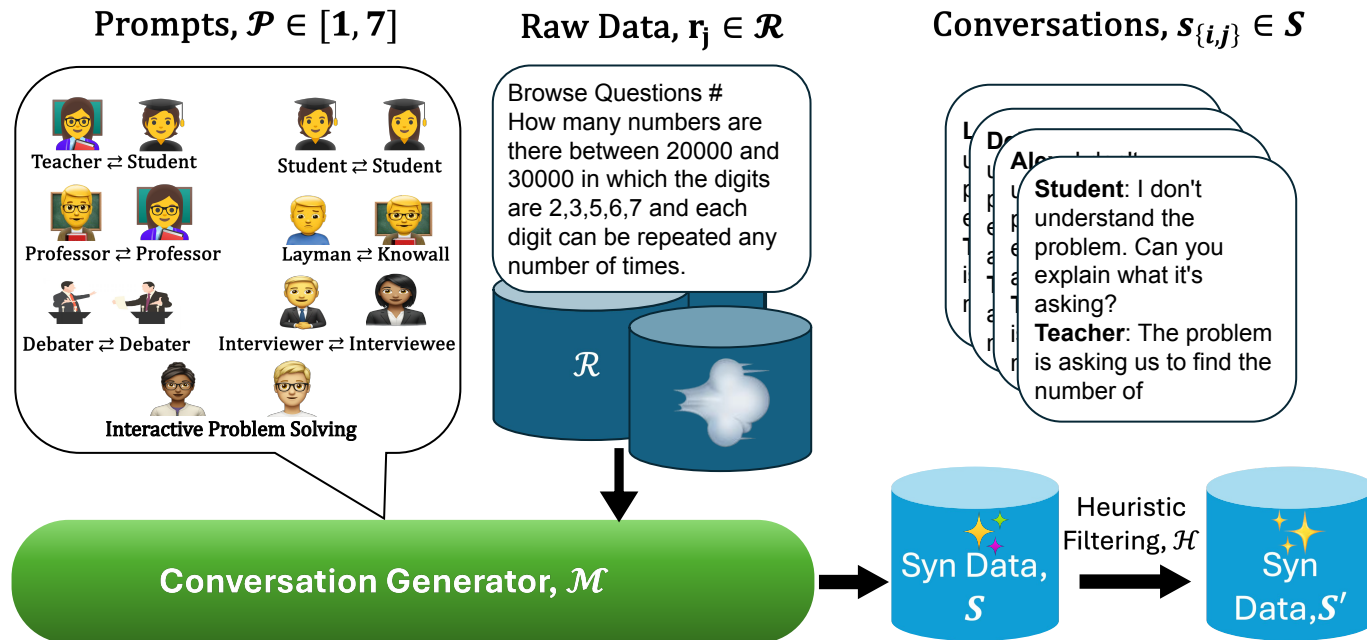
- Data Config (\mathcal{R}):
 - Open Web Math (14B Tokens)
 - Mathematical Web Pages from Common Crawl
- LLM Config (\mathcal{M}):
 - LLaMA3-70B-Instruct
- Generation details
 - Input context length: 500 tokens



With increasing context length the generated conversation length decreases!

How does MIND work?

Proposed Framework



Generate Synthetic Dialogue Corpora using MIND

$\mathcal{R} \rightarrow \text{MIND} \rightarrow \mathcal{S}'$



Generated Dialogues

Methodology

Prompt

Convert the context above as a multi-turn discussions between a teacher and a student. The student has questions about the context and the teacher solves each of them step-by-step.

Raw Text

How many numbers are there between 20000 and 30000 in which the digits are 2,3,5,6,7 and each digit can be repeated any number of times. Since the numbers are > 20000 and < 30000 1st place can be only 2. Remaining 4 places can be occupied by any of the 5 digits. Therefore, Required no. of numbers = $5^4=625$

Conversation



I don't understand the problem. Can you explain what it's asking?

The problem is asking us to find the number of numbers between 20000 and 30000 that can be formed using the digits 2, 3, 5, 6, and 7. Each digit can be repeated any number of times.



Okay, got it. So, why is the first place limited to only 2?

That's because the numbers are greater than 20000 and less than 30000. *If the first digit were 1, the number would be less than 20000, and if it were 3, the number would be greater than 30000. So, the first place can only be 2.*



I see. What about the remaining places? Can they be occupied by any of the given digits?

Yes, that's correct. The remaining 4 places can be occupied by any of the 5 digits: 2, 3, 5, 6, and 7.



Okay, so how do we find the total number of numbers that can be formed?

Since each of the remaining 4 places can be occupied by any of the 5 digits, we have 5 options for each place. To find the total number of numbers, we raise 5 to the power of 4, which is $5^4 = 625$.

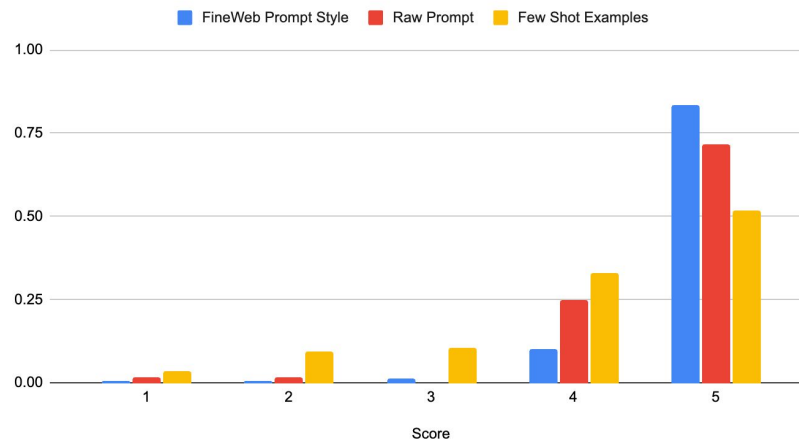


Quality Assessment of Dialogues

Methodology

- Heuristic Analysis
 - If generation length < 50 tokens, discard the conversation
- LLM based scoring
 - Used LLaMA3-70B-instruct to evaluate the quality of the 140k synthetic conversations
 - The score distribution is very skewed
- Qualitative analysis
 - Manually annotated 20 examples
 - Human and LLM rating do not overlap

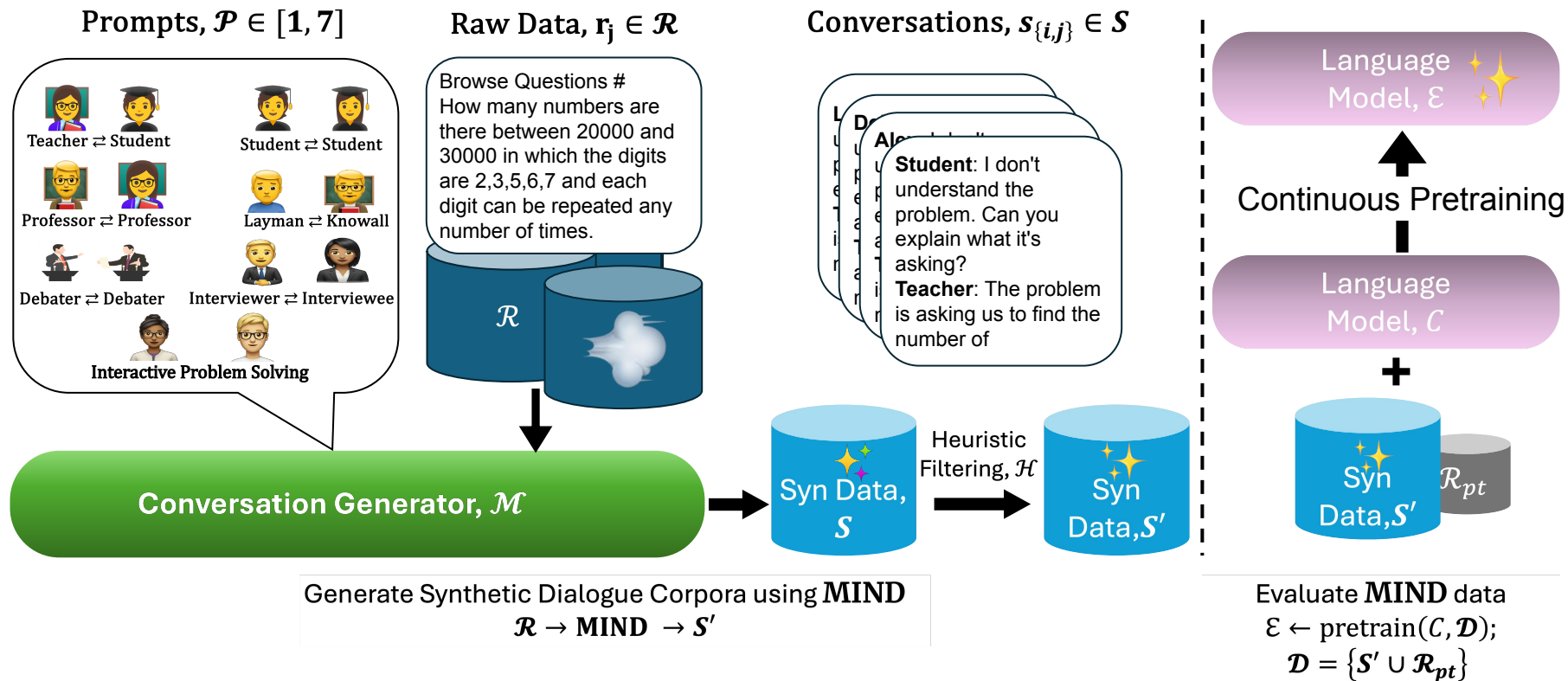
FineWeb Prompt Style vs Raw Prompt vs Few Shot Examples



LLM tends to rate its generation higher most of the times!

How does MIND work?

Proposed Framework



Pretraining Model Configuration

Experimental Setup

- **Base Model:**
 - 7B LLM trained on 700B tokens
 - Pretraining data consists of 13 Common Crawl (CC) corpora alongside other datasets
- **Baseline Model:**
 - Continuously train the base model on 50B additional tokens with a blend that maintains 2:1 ratio of OpenWebMath (OWM) and 13 CC data
- **Experimenting with Synthetic Data:**
 - Follow the same training configuration as the baseline model; only replace raw OWM with Synthetic OWM



Evaluation Metric Details

Experimental Setup

- **Math Benchmarks**

- GSM8k: 8-shot and majority@8
- MATH: 4-shot and pass@1

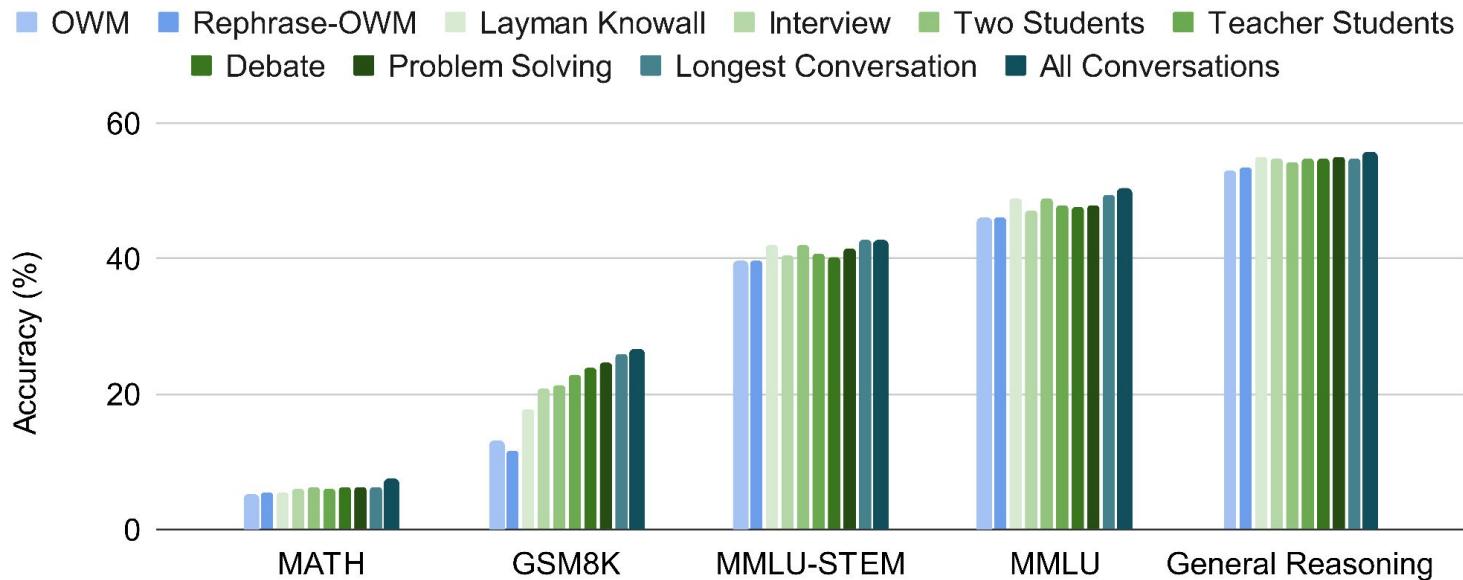
- **Reasoning Benchmarks**

- 0-shot accuracy: ARC-easy, ARC-challenge, PIQA, SIQA, Hellaswag, Winogrande, OpenBookQA, TruthfulQA, CommonsenseQA, RACE.
 - We report average scores for all these benchmarks
- 5-shot accuracy: MMLU which also includes MMLU-STEM



What is the potential of MIND?

Experimental Results



For all styles, MIND works better than raw data!

Rephrasing has very little effect but combining all dialogues yields the best number!

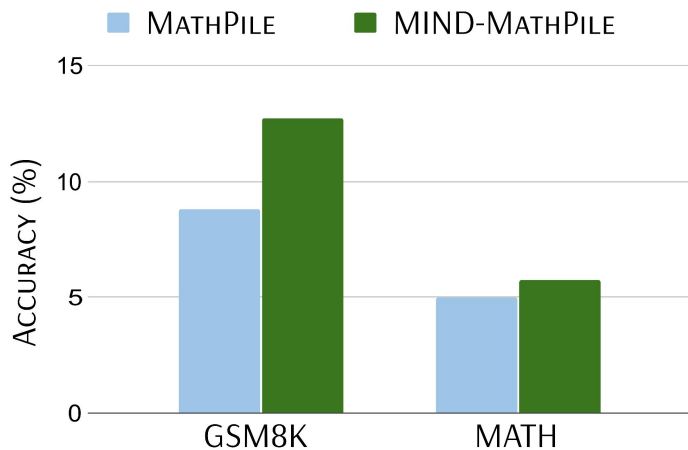


Does MIND scale with data?

Ablation Studies

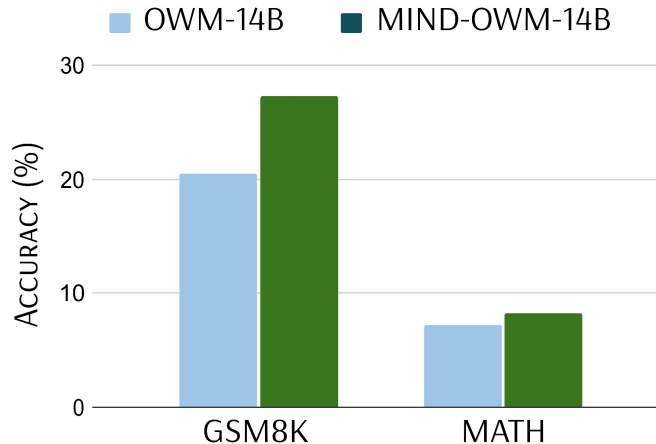
Does MIND work on high-quality datasets?

MATHPILE consists of 9.3B tokens of high-quality data sources such as ArXiv papers, textbooks, and Wikipedia. We see **4% absolute improvement on GSM8k!**



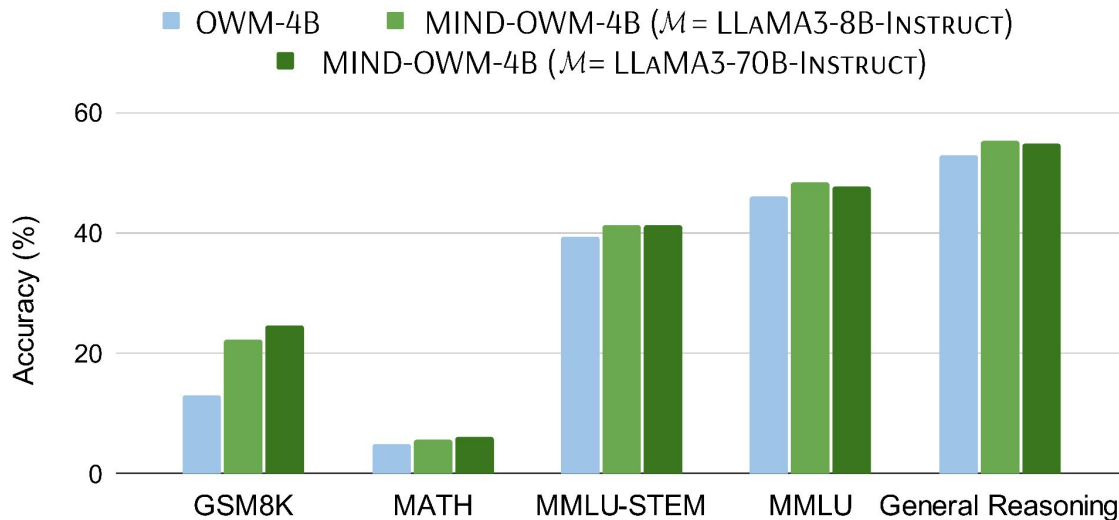
Does MIND scale with data size?

We use the whole OWM 14B token corpus.
7% absolute improvement on GSM8k



Conversation with Weaker LLM?

Ablation Studies



MIND gains are not solely dependent on the capabilities of the larger LLM but are largely driven by the quality and structure of the MIND-generated dialogues.



MIND with Popular LLMs

Ablation Studies

Dataset	Tokens	GSM8K	MATH	MMLU-STEM	MMLU	GENERAL REASONING (Avg)	Avg-Math	Avg-All
DEEPSEEKMATH (Zhihong Shao, 2024)	500B	59.29	4.37	55.41	54.98	55.94	39.69	43.64
MIND-OWM-4B/14B [Combinations*]		57.32	2.36	51.95	56.54	59.16	37.21	43.84

Model	Math Data	Tokens	MATH (lvl 5)	GSM8K	MMLU-STEM	AVG-math	AVG-reasoning
DeepSeekMath	OWM	500B	1.59%	39.42%	49.92%	30.31%	55.47%
DeepSeekMath	MIND-OWM	500B	2.36%	57.32%	51.95%	37.21%	58.90%

The effectiveness of MIND is solid regardless of the quality of the pretrained model!



Does prompt style matter?

Ablation Studies



I've been **thinking about** the formalism for Cooper pairing, and **I believe** it can be easily extended from the case of a single-particle density matrix to a two-particle density matrix.

Ah, yes! That's similar to Yang's approach. But **have you considered** the formalism outlined by Leggett in his book Quantum Liquids? It was first developed by Penrose and Onsager.



Yes, I'm familiar with it. They **conclude** that many properties of Bose-Einstein Condensation can be obtained by examining the diagonalized density matrix.

Does prompt style matter?

Ablation Studies



I've been **thinking about** the formalism for Cooper pairing, and **I believe** it can be easily extended from the case of a single-particle density matrix to a two-particle density matrix.

Ah, yes! That's the formalism outlined in the first development of

ed the
was



**No explanations!
No clarifications!
No reasoning chains because the
assumptions is that both characters
have same knowledge.**

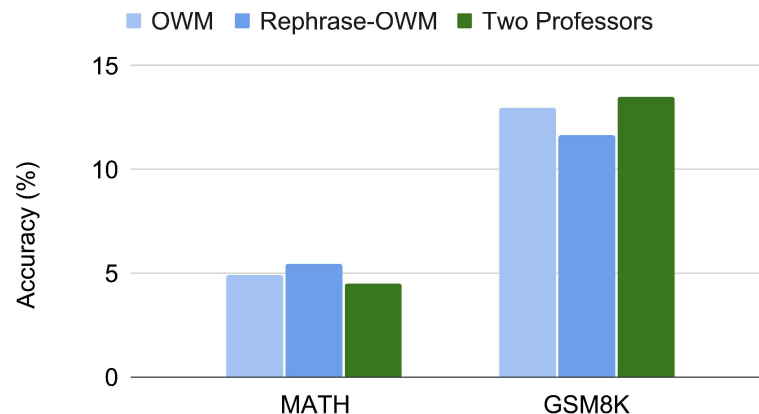
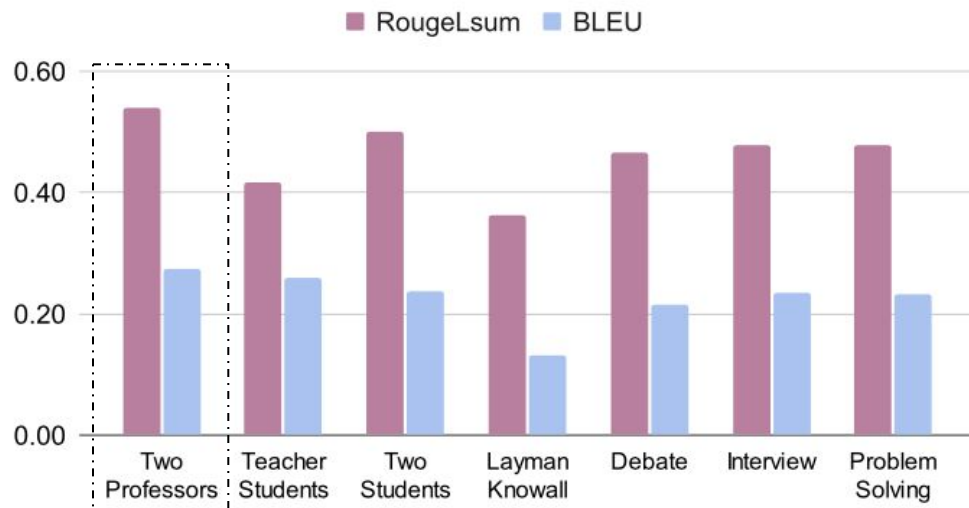


properties of
Bose-Einstein Condensation can be obtained by examining the diagonalized density matrix.



Does prompt style matter?

Ablation Studies



Two Professors style has highest Rouge and BLEU scores with raw text.

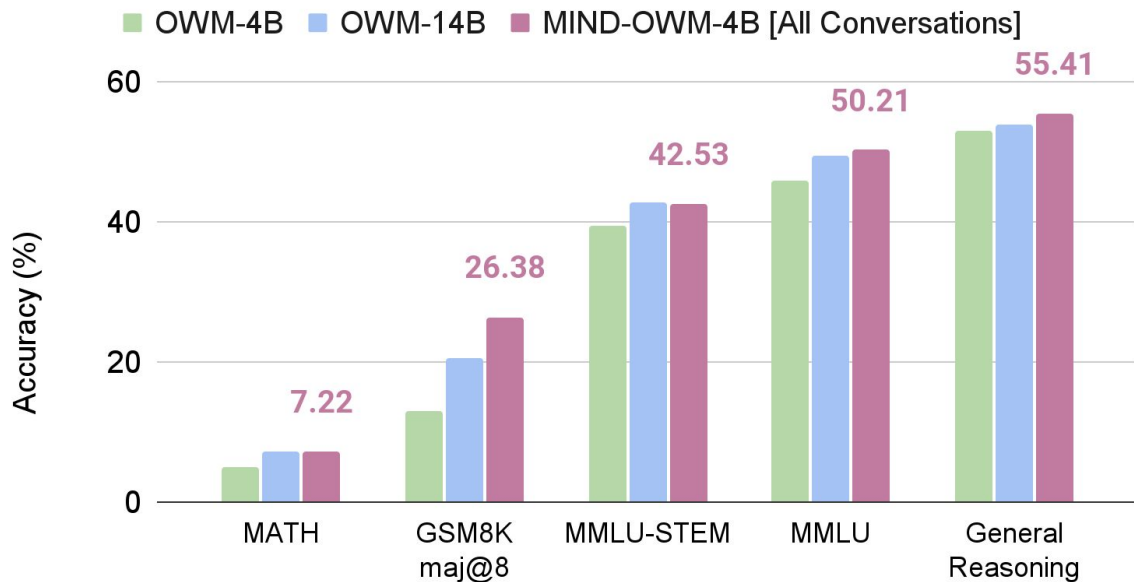
Layman-Knowall and Teacher-Student styles have the lowest similarity with raw text.

Two Professors style does not provide improvement over raw text because on no knowledge gap between the participants!



Potential for Infinite Data

Conclusion

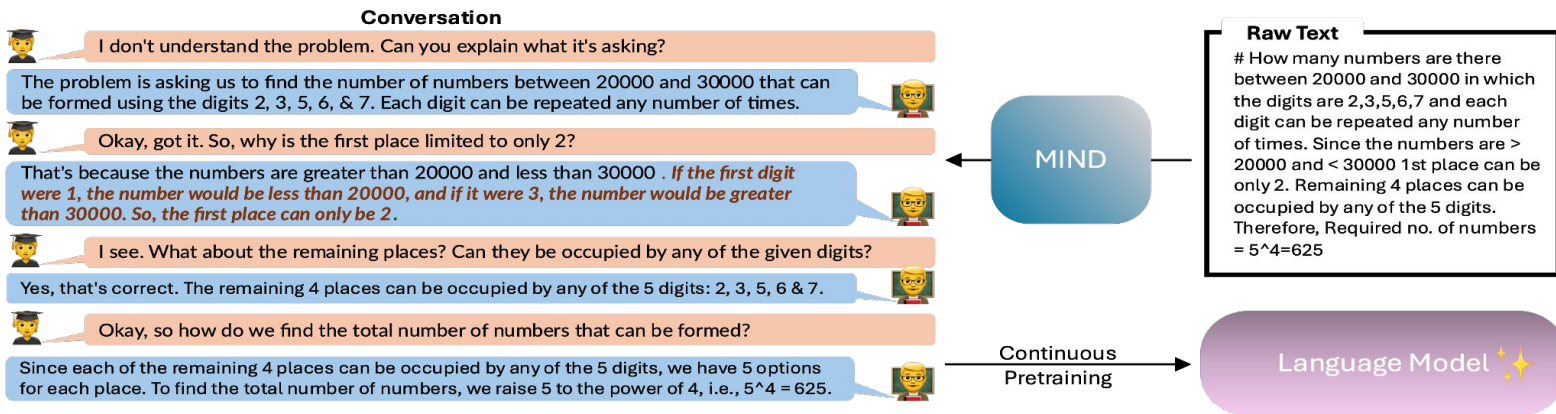


All styles of conversations derived from a small subset of OWM-4B is better than a 3.6× large raw corpus OWM-14B.



Summarizing MIND

Conversations yields Significant Boost across all Reasoning Categories



Boost in Math Reasoning!

