

Datasheet for MathChatSync-reasoning

Anonymous

1 Datasheet

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description. MathChatSync-reasoning was developed to enable supervised training on multi-turn conversations with explicit reasoning for each assistant message. The goal was to address the lack of a publicly available multi-turn reasoning dataset.

Composition

What do the instances that comprise the dataset represent? Are there multiple types of instances? Please describe. Each instance is a multi-turn mathematics tutoring dialogue consisting of human and assistant messages. Each assistant message is augmented with a step-by-step reasoning trace.

How many instances are there in total (of each type, if appropriate)? There are 8,797 conversations in total. Each conversation contains multiple turns, and each assistant turn includes a synthetic reasoning.

Does the dataset contain all possible instances or is it a sample?

If it is a sample, please describe the broader population and representativeness. It is a stratified sample from the original MathChatSync corpus. Depth-balanced sampling was used to mitigate overrepresentation of six-turn dialogues, ensuring a more even distribution of conversation lengths.

What data does each instance consist of? “Raw” data or features?

Please provide a description. Each instance includes the raw text of the human prompt, the assistant’s response, and a synthetically generated reasoning trace. No additional high-level features are provided.

Is there a label or target associated with each instance? If so, please describe.

There is no explicit label. The main “target” is the assistant’s full response plus its accompanying reasoning tokens.

Are relationships between individual instances made explicit? If so, please describe how.

No. Each conversation is isolated; there are no explicit links or relationships between different dialogues.

Are there recommended data splits (e.g., training, validation, testing)?

If so, please describe. No official splits are included. Users may randomly partition into training, validation, and test sets as needed because conversations are generally independent.

Is the dataset self-contained, or does it rely on external resources?

If it links to external resources, please describe. It is self-contained, with all augmented dialogues available in the dataset files. No external resources are required.

Does the dataset contain data that might be considered confidential?

If so, please describe. No. It is synthetic math content with no inclusion of private or personal data, so it does not include confidential information.

Does the dataset contain data that might be offensive or cause anxiety? If so, please describe why. Not

to our knowledge. The content is focused on math problems and related tutoring messages.

Does the dataset relate to people?

No, it does not contain personal data or real individuals.

Collection Process

How was the data associated with each instance acquired? Please describe.

It was derived from the MathChatsync corpus, which was synthetically generated. Then, GPT-4.1-mini was used to generate reasoning texts for each assistant turn. Each final instance pairs the original user-assistant conversation with the newly added reasoning traces.

What mechanisms or procedures were used to collect the data? How were these validated?

Original dialogues came from synthetic generation in MathChatsync. The augmentation with reasoning used GPT-4.1-mini via a controlled prompt to output hidden rationales. Basic quality checks ensured the reasoning aligned with the assistant’s final answers.

If the dataset is a sample from a larger set, what was the sampling strategy?

A depth-balanced sampling: The dataset reduces over-represented six-turn dialogues and ensures coverage across varied conversation lengths.

Who was involved in the data collection process and how were they compensated?

Data is synthetic; no direct human participants or crowdworkers were employed beyond the dataset builders and maintainers.

Over what timeframe was the data collected?

Does this timeframe match the creation timeframe of the data? The original MathChatsync was developed earlier in 2024. The reasoning augmentation was performed in 2025 for the release of MathChatSync-reasoning.

Were any ethical review processes conducted? If so, describe.

Not applicable; the content is synthetic math dialogue with no direct human subjects.

Does the dataset relate to people?

If not, you may skip the remaining questions in this section. No. It is a synthetic tutoring scenario involving no real individuals.

Did you collect the data from the individuals in question directly, or

obtain it via third parties? Not applicable. The data is synthetic.

Were the individuals in question notified about the data collection? If so, please describe. Not applicable.

Did the individuals in question consent to the collection and use of their data? If so, how was consent obtained? Not applicable.

If consent was obtained, were they provided a mechanism to revoke consent? If so, please describe. Not applicable.

Has an analysis of the potential impact of the dataset on data subjects been conducted? If so, describe. Not applicable.

Preprocessing/cleaning/labeling

Was any cleaning/labeling of the data done? Please describe. Yes. The main “labeling” step was adding reasoning tokens. The original text remains in place, with minimal cleanup beyond verifying basic alignment between the final answer and the synthetic reasoning.

Was the “raw” data saved in addition to the preprocessed data? If so, provide a link. MathChatSync is publicly available. The augmented version includes both original dialogues and reasoning tokens in the same file.

Is the software used to preprocess the instances available? If so, provide a link. A script for augmenting the original dataset with GPT-4.1-mini is included. The data generation

code is publicly accessible in the associated anonymous repository.

Uses

Has the dataset been used for any tasks already? If so, describe. Yes, it has been used to fine-tune language models for mathematical reasoning and tutoring tasks, as described in the associated paper.

Is there a repository linking to papers or systems that use the dataset? If so, provide a link. Users can refer to the dataset’s Hugging Face repository and the accompanying arXiv submission for references to ongoing projects using MathChatSync-reasoning.

What (other) tasks could the dataset be used for? It can be used to train multi-turn dialogue agents, study step-by-step problem-solving, design math tutoring systems, or develop new data augmentation strategies for reasoning.

Is there anything about the composition or collection process that might impact future uses? If so, please describe. Because all dialogues and reasoning are synthetic, some nuance of real student-teacher interactions might be missing, potentially affecting real-world fidelity.

Are there tasks for which the dataset should not be used? If so, please describe. It is not recommended for sensitive or personal user modeling, since it does not reflect real user data.