Supplementary Materials for NeurIPS 2024 Submitted Paper: FormulaReasoning: A Dataset for Formula-Based Numerical Reasoning

Xiao Li Bolin Zhu Sichen Liu Yin Zhu Yiwei Liu Gong Cheng*
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
{xiaoli.nju,bolinzhu,sichenliu,yinzhu,ywliu}@smail.nju.edu.cn
gcheng@nju.edu.cn

1 Dataset Card

1.1 Dataset documentation and intended uses

1.1.1 Motivation

1. 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.

The motivation behind constructing FormulaReasoning comes from the need to address the limitations of existing numerical reasoning datasets. While numerical reasoning has seen significant advancements with the rise of LLMs and specialized datasets, current datasets often lack knowledge-guided reasoning process. They typically rely on implicit commonsense knowledge rather than explicit formulas, which becomes problematic when LLMs encounter hallucinations.

To overcome these limitations, FormulaReasoning was created to emphasize the use of specific formulas in numerical reasoning. Unlike previous datasets that primarily rely on implicit knowledge, FormulaReasoning requires explicit formula-based reasoning. This shift introduces a higher level of challenge and reflects real-world numerical problem-solving scenarios better.

2. Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

FormulaReasoning is created by Xiao Li, Bolin Zhu, Sichen Liu, Yin Zhu, Yiwei Liu and Gong Cheng from the State Key Laboratory for Novel Software Technology, Nanjing University.

3. Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

This work was supported by the CIPSC-SMP-Zhipu.AI Large Model Cross-Disciplinary Fund.

1.1.2 Composition

1. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

*Corresponding author

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

The data within the dataset exclusively comprises elementary physics questions based on daily life scenarios, all organized in text format, without photos, specific people information or specific countries.

2. How many instances are there in total (of each type, if appropriate)?

We divided FormulaReasoning into training, *id* (in-distribution) test, and *ood* (out-of-distribution) test, comprising 4,608, 421 and 391 questions, respectively.

3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

FormulaReasoning is not from a larger set.

4. What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Each instance consists of a question, the formulas, the parameters within these formulas and their corresponding numerical values, textual explanations, and the final numerical answer. See https://github.com/nju-websoft/FormulaReasoning for more details.

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

Yes, each instance contains textual explanations, and the final numerical answer.

6. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

No.

7. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

N/A.

8. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Yes. We divided FormulaReasoning into training, *id* (in-distribution) test, and *ood* (out-of-distribution) test, comprising 4,608, 421 and 391 questions, respectively. We required that all formulas in the id test must appear in the training set, whereas in the ood test, each question involves at least one formula that has not been seen in the training set. This division is designed to evaluate the generalization capabilities of fine-tuned models on formulas that they have not previously encountered.

9. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Currently, there are no known errors, noise, or redundancies. We have addressed these occurrences during the annotation process.

10. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

Yes, FormulaReasoning is self-contained, and it doesn't rely on any external resources.

11. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

No.

12. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No. Firstly, it is unlikely for harmful information to appear in the questions designed for middle school education. Secondly, we have not identified such information within the dataset.

13. Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No.

1.1.3 Collection Process

1. How was the data associated with each instance acquired?

See Section 3 in the main paper.

2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)?

See Section 3 in the main paper.

3. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Our FormulaReasoning is not sampled from a larger set.

4. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

A total of 5 graduate students participated in the annotation work, and 108 high school students were involved in the human performance tasks. For more details, see Section 3 and Section 4 in the main paper.

5. Over what timeframe was the data collected?

The questions in FormulaReasoning were derived from junior high school physics examinations in China over the past 14 years (2010 – 2024).

6. Were any ethical review processes conducted (e.g., by an institutional review board)?

The ethical review board of our department has approved our experiment.

1.1.4 Preprocessing/cleaning/labeling

1. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?

Yes. For more details, see Section 3 in the main paper.

2. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

Yes, the raw data has been included in the released dataset.

3. Is the software that was used to preprocess/clean/label the data available?

Yes, they are includes in our GitHub repository.

1.1.5 Uses

1. Has the dataset been used for any tasks already? If so, please provide a description.

Yes, in this paper, we utilized the dataset to evaluate the reasoning ability of language models.

2. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

N/A. Currently, there have been no external works that have utilized FormulaReasoning.

3. What (other) tasks could the dataset be used for?

FormulaReasoning can be utilized for evaluating the reasoning ability of language models, particularly in scenarios requiring knowledge (formulas). Additionally, the formula database we constructed can be employed for evaluating retrieval-augmented generation models. Furthermore, we partitioned the test set into id and ood tests for assessing the generalization ability of language models.

4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

No. Our data originates from elementary physics questions based on everyday life scenarios, excluding any potentially harmful information.

5. Are there tasks for which the dataset should not be used? If so, please provide a description.

No.

1.1.6 Distribution

1. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

No. We only open source the datasets through public channels: https://github.com/nju-websoft/FormulaReasoning.

2. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

Our code is available at `https://github.com/nju-websoft/FormulaReasoning` under the Apache 2.0 License.

Our data is available at `https://zenodo.org/doi/10.5281/zenodo.11408109` under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

DOI: 10.5281/zenodo.11408109.

Croissant metadata: `https://huggingface.co/api/datasets/xli/FormulaReasoning/croissant`.

3. When will the dataset be distributed?

We have distributed FormulaReasoning.

4. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Our code is distributed under the Apache License, Version 2.0. Our data is distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

5. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

6. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

1.1.7 Maintenance

1. Who will be supporting/hosting/maintaining the dataset?

The Authors.

2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Contact authors via emails listed under the title or through GitHub issues.

3. Is there an erratum? If so, please provide a link or other access point.

No.

4. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (e.g., mailing list, GitHub)?

Updates, if any, will be provided on GitHub by the authors.

5. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No, FormulaReasoning doesn't relate to people.

6. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.

N/A.

7. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.

1.2 Accessibility

Our code is available at `https://github.com/nju-websoft/FormulaReasoning` under the Apache 2.0 License.

Our data is available at `https://zenodo.org/doi/10.5281/zenodo.11408109` under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

DOI: 10.5281/zenodo.11408109.

1.3 Croissant Metadata

Croissant metadata: `https://huggingface.co/api/datasets/xli/FormulaReasoning/croissant`.

1.4 Author Statement

We state that we bear all responsibility in case of violation of rights, etc., and confirmation of the data license.

1.5 Hosting, Licensing, and Maintenance Plan.

Our code is distributed under the Apache License, Version 2.0. Our data is distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.

We will maintain the dataset on GitHub and Zenodo, and promptly update FormulaReasoning on GitHub and Zenodo in the event of any updates.