OlympicArena datacard

OlympicArena

OlympicArena is a comprehensive, highly-challenging, and rigorously curated benchmark featuring a detailed, fine-grained evaluation mechanism designed to assess advanced AI capabilities across a broad spectrum of Olympic-level challenges.

Dataset Link

Github Repo

Hugging Face Dataset

Croissant metadata

Data Card Author(s)

- Zhen Huang

Dataset Owners

Team(s)

Generative AI Research Lab (GAIR)

Contact Detail(s)

- **Dataset Owner(s):** Generative AI Research Lab (GAIR)
- **Affiliation:** Shanghai Jiao Tong University
- **Group Email:** gair.olympicarena@gmail.com
- **Website:** https://gair-nlp.github.io/OlympicArena/

Author(s)

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, Yikai Zhang, Yuqing Yang, Ting

Wu, Binjie Wang, Shichao Sun, Yang Xiao, Yiyuan Li, Fan Zhou, Steffi Chern, Yiwei Qin, Yan Ma, Jiadi Su, Yixiu Liu, Yuxiang Zheng, Shaoting Zhang, Dahua Lin, Yu Qiao, Pengfei Liu

Funding Sources

Institution(s)

- Shanghai Jiao Tong University

Funding or Grant Summary(ies)

N/A

Dataset Overview

Data Subject(s)

- Data about scientific problems

Dataset Snapshot

Category	Data
Size of Dataset	266MB
Number of Instances	11163
Number of Fields	14
Labeled Classes	N/A
Number of Labels	N/A
Average Labels Per Instance	N/A
Algorithmic Labels	N/A
Human Labels	N/A

Content Description

This benchmark encompasses seven disciplines: Mathematics, Physics, Chemistry, Biology, Geography, Astronomy, and Computer Science. Each discipline is divided into two splits: validation (val) and test. The validation split

includes publicly available answers for small-scale testing and evaluation, while the test split does not disclose the answers, users could submit their results through our platform.

Sensitivity of Data

N/A

Dataset Version and Maintenance

Maintenance Status

Actively Maintained - No new versions will be made available, but this dataset will be actively maintained, including but not limited to updates to the data.

Version Details

Current Version: 1.0

Last Updated: 06/2024

Release Date: 06/2024

Maintenance Plan

Versioning: N/A.

Feedback: For feedback, reach out to gair.olympicarena@gmail.com or open an issue on our Github Repo

Example of Data Points

Primary Data Modality

- Text
- Image

Sampling of Data Points

Explore OlympicArena on our Hugging Face Dataset.

Data Fields

Field Name	Description
id	The unique identifier for each problem
problem	The problem statement
prompt	The prompt used as input to the model (as used in the paper); we also encourage users to try their own prompts
figure_urls	Links to images that appear in the problem, in order
answer	The answer to the problem
answer_type	The type of the answer
unit	The unit corresponding to the answer
answer_sequence	The sequence in which the model should provide answers if multiple quantities are required
type_sequence	The sequence of answer_type for each quantity if multiple quantities are required
test_cases	Test cases used for evaluation in CS code generation problems
subject	The subject of the problem
language	The language of the problem, where EN represents English and ZH represents Chinese
modality	The modality type of the problem statement, where text-only indicates the problem statement does not contain images, and multi-modal indicates the problem statement contains images

Typical Data Point

Below is a dev example from Math:

```
# Problem
If $x$ is a real number so $3^{x}=27 x$, compute $\log _{3}\l
eft(\frac{3^{3^{x}}}{x^{3}}\right)$.

# Answer
81

# Solution
```

```
We plug in the condition that we were given initially to get a value of $\log _{3}\left(\frac{3^{27 x}}{x^{27}}\right)$. We can simplify this by using the equality again to get $\log _{3}\left(\frac{\left(3^{x}\right)^{27}}{x^{27}}\right)=\log _{3}\left(\frac{27^{27} * x^{27}}{x^{27}}\right)=\log _{3}\left(27^{27}\right)=$ $27 * 3=81$.
```

Atypical Data Point

The dataset does not contain atypical data points as far as we know.

Motivation & Intentions

Motivations

Purpose(s)

- research

Domain(s) of Application

Machine Learning, Natural Language Processing

Motivating Factor(s)

- Advancing AI towards superintelligence, equipping it to address more complex challenges in science and beyond.

Intended Use

Dataset Use(s)

- Safe for research use

Suitable Use Case(s)

- Search for better hyperparameter during training. For example, determine the optimal pre-train data
  mixing scheme.

Unsuitable Use Case(s)

- The dataset is created for model evaluation. It is not intended to be used as pre-training data.

Provenance

Collection

Method(s) Used

- human annotation
- website parsing
- Scraped
- AI annotated

Source Description(s)

- The primary source is Olympic exams freely available on the internet.

Collection Cadence

Static: Data was collected once from single or multiple sources.

Data Processing

- All questions are parsed by tools and annotated by human annotators.
- Convert complex mathematical notations into standard LATEX formats.
- Deduplication and cleaning
- Human validation

Use in ML or AI Systems

Dataset Uses(s)

- Testing
- Validation

- Development or Production Use

Usage Guidelines(s)

Please visit our Github Repo for detailed information.

Distribution(s)

Set	Number of data points
Val	638
Test	9977
OT	548

Author statement

We recognize that our benchmark may incorporate data from sources where explicit consent has not been obtained from each individual contributor. We have implemented comprehensive measures to ensure the ethical utilization of data and adherence to intellectual property rights. Nonetheless, we accept full responsibility for any potential violations of rights or licensing issues that may emerge from the dataset.

Licenses

License CC BY-NC-SA 4.0