C-EVAL

C-Eval is a comprehensive Chinese evaluation suite designed to assess advanced knowledge and reasoning abilities of foundation models. It consists of 13948 multi-choice questions spanning 52 diverse disciplines and four difficulty levels.

Dataset Link

Github Repo

Official Website

Data Card Author(s)

- Yuzhen Huang

Dataset Owners

Team(s)

C-EVAL Team

Contact Detail(s)

- **Dataset Owner(s):** C-EVAL Team
- **Affiliation:** Shanghai Jiao Tong University
- **Group Email:** ceval.benchmark@gmail.com
- **Website:** https://cevalbenchmark.com/

Author(s)

- Yuzhen Huang
- Yuzhuo Bai
- Zhihao Zhu
- Junlei Zhang
- Jinghan Zhang
- Tangjun Su
- Junteng Liu
- Chuancheng Lv
- Yikai Zhang
- Jiayi Lei
- Yao Fu

- Maosong Sun
- Junxian He

Funding Sources

Institution(s)

- Shanghai Jiao Tong University

Funding or Grant Summary(ies)

N/A

Dataset Overview

Data Subject(s)

- Data about natural phenomena
- Data about places and objects

Dataset Snapshot

Category	Data
Size of Dataset	3.5 MB
Number of Instances	13948
Number of Fields	8
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

Each datapoint in the dataset contains one multiple-choice question with four choices and its corresponding label. We randomly split datapoints into a development set, a validation set, and a test set within each subject. For development set, we provide explanations for each question. And we keep the labels of test set private to ensure the fair use of C-EVAL.

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: 06/2023

Release Date: 05/2023

Maintenance Plan

Versioning: N/A. C-EVAL is a static dataset. Minor releases correspond to any errors fixed in the dataset.

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

Example of Data Points

Primary Data Modality

- Text Data

Sampling of Data Points

Explore C-EVAL on our official website.

Data Fields

Field Name	Field Value	Description
id	Integer	the order number of the data point
question	string	the main body of multiple-choice question
A	string	choice A for the question
B	string	choice B for the question
C	string	choice C for the question
D	string	choice D for the question
answer	string	the ground-truth answer of the question[1]
explanation	string	the explanation of the question[2]

[1] For test set, we keep the answer private.

[2] Only the questions in the development set contain explanations.

Typical Data Point

Below is a dev example from computer network:

```
id: 1
question: 25 °C时，将pH=2的强酸溶液与pH=13的强碱溶液混合，所得混合液的pH=11，则强酸溶液与强碱溶液 的体积比是(忽略混合后溶液的体积变化)____
A: 11:1
B: 9:1
C: 1:11
D: 1:9
answer: B
explantion:
1. pH=13的强碱溶液中c(OH-)=0.1mol/L，pH=2的强酸溶液中c(H+)=0.01mol/L，酸碱混合后pH=11，即c(OH-)=0.001mol/L。
2. 设强酸和强碱溶液的体积分别为x和y，则：c(OH-)=(0.1y-0.01x)/(x+y)=0.001，解得x:y=9:1。
```

Atypical Data Point

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

Motivations & Intentions

Motivations

Purpose(s)

- Research

Domain(s) of Application

Machine Learning, Natural Language Processing

Motivating Factor(s)

- Evaluating the advanced abilities of foundation models in a Chinese context, one of the most widely spoken language in the world.
- Narrow the gap between Chinese large language model development and evaluation.

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.
- Assess advanced knowledge and reasoning abilities of foundation models in a Chinese context.

Unsuitable Use Case(s)

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

Citation Guidelines

BiBTeX:

```
@article{huang2023ceval,
title={C-Eval: A Multi-Level Multi-Discipline Chinese Evaluation Suite for Foundation Models},
author={Huang, Yuzhen and Bai, Yuzhuo and Zhu, Zhihao and Zhang, Junlei and Zhang, Jinghan and Su,
Tangjun and Liu, Junteng and Lv, Chuancheng and Zhang, Yikai and Lei, Jiayi and Fu, Yao and Sun,
Maosong and He, Junxian},
journal={arXiv preprint arXiv:2305.08322},
year={2023}
}
```

Provenance

Collection

Method(s) Used

- Artificially Generated
- Scraped or Crawled
- Manually parsed

Source Description(s)

- The primary source is mock exams freely available on the internet.
- A portion of the college-level questions are past exam questions from top universities in China, publicly shared by the students.
- A minor fraction of college questions are mock questions for the national graduate entrance exam, sourced from the Weipu website[1]. We have obtained their authorization to include around 2000 such questions into C-Eval.

[1] https://kaoyan.cqvip.com/view/postgraduate/index.aspx

Collection Cadence

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

Data Processing

- All questions are subsequently parsed – automatically when possible, and otherwise manually by the authors – into a structured format.
- Convert complex mathematical notations into standard LATEX formats.
- Deduplication and cleaning
- Human validation

Use in ML or AI Systems

Dataset Use(s)

- Testing
- Validation
- Development or Production Use

Usage Guideline(s)

Please visit our Github Repo for detailed information.

Distribution(s)

Set	Number of data points
Dev	260
Valid	1346
Test	12342

Licenses