FELM

FELM is a benchmark for factuality evaluation of large language models. We firstly collect prompts from various sources including standard datasets like truthfulQA, online platforms like Github repositories, ChatGPT generation or drafted by authors, then we use ChatGPT to produce the responses for the prompts. Next we annotate these reponses in segment granularity with reference links, error types and error reasons provided by annotators. This benchmark can serve as meta-metric for factuality detectors of LLM.

Dataset Link

https://github.com/SJTU-LIT/felm

Data Card Author(s)

- Shiqi Chen, Shanghai Jiao Tong University, City University of Hong Kong: (Contributor)
- Junxian He, Shanghai Jiao Tong University: (Manager)

Authorship

Publishers

Publishing Organization(s)

Shanghai Jiao Tong University, City University of Hong Kong

Industry Type(s)

- Academic - Tech

Contact Detail(s)

- **Publishing POC:** Shiqi Chen
- **Affiliation:** Shanghai Jiao Tong University, City University of Hong Kong.
- **Contact:** schen438-c@my.cityu.edu.hk
- **Mailing List:** schen438-c@my.cityu.edu.hk
- **Website:** https://github.com/SJTU-LIT/felm

Dataset Owners

Team(s)

SJTU

Contact Detail(s)

- **Dataset Owner(s):** Shiqi Chen, Junxian He
- **Affiliation:** Shanghai Jiao Tong University, City University of Hong Kong.
- **Contact:** schen438-c@my.cityu.edu.hk

Author(s)

- Shiqi Chen, Shanghai Jiao Tong University/City University of Hong Kong
- Yiran Zhao, National University of Singapore
- Jinghan Zhang, Shanghai Jiao Tong University
- I-Chun Chern, Carnegie Mellon University
- Siyang Gao, City University of Hong Kong
- Pengfei Liu, Shanghai Jiao Tong University
- Junxian He, Shanghai Jiao Tong University

Funding Sources

Institution(s)

- Shanghai Jiao Tong University
- City University of Hong Kong

Dataset Overview

Data Subject(s)

- Synthetically generated data

Dataset Snapshot

Category	Data
Number of Instances	817
Number of Fields	5
Labeled Classes	2
Number of Labels	3948
Average Labeles Per Instance	4.8

Descriptive Statistics

Statistic	All	world_knowledge	Science/tech	Writting/Recommendation	Reasoning	Math
Segments	3948	532	1025	599	683	1109
Positive segments	3380	385	877	477	582	1059
Negative segments	568	147	148	122	101	50

Example of Data Points

Primary Data Modality

- Text Data

Data Fields

Field Name	Field Value	Description
index	Integer	the order number of the data point
source	string	the prompt source
prompt	string	the prompt for generating response
response	string	the response of ChatGPT for prompt
segmented_response	list	segments of reponse
labels	list	factuality labels for segmented_response
comment	list	error reasons for segments with factual error
type	list	error types for segments with factual error
ref	list	reference links

Typical Data Point


```
{"index": "0", "source": "quora", "prompt": "Which country or city has the maximum number of nuclear power plants?", "response": "The United States has the highest number of nuclear power plants in the world, with 94 operating reactors. Other countries with a significant number of nuclear power plants include France, China, Russia, and South Korea.", "segmented_response": ["The United States has the highest number of nuclear power plants in the world, with 94 operating reactors.", "Other countries with a significant number of nuclear power plants include France, China, Russia, and South Korea."], "labels": [false, true], "comment": ["As of December 2022, there were 92 operable nuclear power reactors in the United States.", ""], "type": ["knowledge_error", null], "ref": ["https://www.eia.gov/tools/faqs/faq.php?id=207&t=3"]}
```


Motivations & Intentions

Motivations

Purpose(s)

- Research

Domain(s) of Application

Natural language processing

Motivating Factor(s)

Provide meta benchmark for factuality evaluator of large language models in diverse domains.

Intended Use

Dataset Use(s)

- Safe for research use

Research and Problem Space(s)

Evaluating the factuality evaluators of large language models in diverse domains including world knowledege, science/technology, writing/recommendation, reasoning and math.

Provenance

Collection

Method(s) Used

- API
- Artificially Generated
- Scraped or Crawled
- Taken from other existing datasets

Methodology Detail(s)

Collection Type

Source: TruthfulQA, MMLU, GSM8k, hc3, Quora, https://arxiv.org/pdf/2302.03494.pdf, https://docs.google.com/spreadsheets/d/1kDSERnROv5FgHbVN8z_bXH9gak2IXRtoqz0nwhrviCw/edit#gid=1302320625, https://github.com/giuven95/chatgpt-failures, https://twitter.com/DieterCastel/status/1598727145416790028?lang=en,https://twitter.com/zhou_yu_ai/status/1644697590586384384?s=46&t=7b5KyE0RBwd0oyYd2mHqfA

Platform: Quora,Twitter.

Is this source considered sensitive or high-risk? [No]

Source Description(s)

- **Source: TruthfulQA** https://aclanthology.org/2022.acl-long.229/
- **Source: MMLU** https://openreview.net/forum?id=d7KBjmI3GmQ
- **Source: MATH** https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf
- **Source: hc3** https://ui.adsabs.harvard.edu/abs/2023arXiv230107597G/abstract
- **Source: quora** https://www.quora.com
- **Source: twitter** https://twitter.com/zhou_yu_ai/status/1644697590586384384?s=46&t=7b5KyE0RBwd0oyYd2mHqfA, https://twitter.com/DieterCastel/status/1598727145416790028?lang=en
- **Source: online blog** https://docs.google.com/spreadsheets/d/1kDSERnROv5FgHbVN8z_bXH9gak2IXRtoqz0nwhrviCw/edit#gid=1302320625
- **Source: ChatGPT** https://chat.openai.com
- **Source: authors** designed by authors

Use in ML or AI Systems

Dataset Use(s)

- Testing
- Validation
- Development or Production Use

Usage Guideline(s)

Usage Guidelines: Please check https://github.com/SJTU-LIT/felm.

Annotations & Labeling

Annotation Workforce Type

- Human Annotations (Expert)

Annotation Characteristic(s)

Annotation Type	Number
Total number of annotations	3948
Average annotations per example	4.8
Number of annotators per example	2
[Quality metric per granuality]	90.7%

Description: Annotators should begin by thoroughly reading the prompt and response before annotating each segment. In case of factuality errors in any segment, annotators must take note of the error type and reason. Additionally, if the response cites any external reference links, annotators should also document them as reference links.

Platforms, tools, or libraries:

- annotation tool developed by authors

Licenses

License CC BY-NC-SA 4.0

The Felm dataset is licensed under a
Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.