Datasets: lmsys/lmsys-chat-1m

Modalities: Text | Formats: parquet | Size: 1M - 10M | ArXiv: arxiv:2309.11998 | Libraries: Datasets | Dask | Croissant + 1

Dataset Viewer

Auto-converted to Parquet | API | Full Screen Viewer

Split (1)
train · 1M rows

Search is not available for this dataset | SQL Console

conversation_id string · *lengths* 32 32	model string · *classes* 25 values	conversation list · *lengths* 2 430
33f01939a744455c869cb234afca47f1	wizardlm-13b	[{ "content": "how can identity services help protect me against
1e230e55efea4edab95db9cb87f6a9cb	vicuna-13b	[{ "content": "Beside OFAC's sel that target the listed individiua
0f623736051f4a48a506fd5933563cfd	vicuna-13b	[{ "content": "You are the text and you must complete the assista
e5c923a7fa3f4893beb432b4a06ef222	palm-2	[{ "content": "The sum of the pe equal squares is 36 cm. Find the
8ad66650dced4b728de1d14bb04657c1	vicuna-13b	[{ "content": "What is the type in the following code defined as
aa041ed88edd4100bde61b8d68fc7288	wizardlm-13b	[{ "content": "I have 1000 docun from a website. So as not to ove

Downloads last month 2,358

Use this dataset | Edit dataset card

Size of downloaded dataset files: 1.49 GB

Size of the auto-converted Parquet files: 1.49 GB

Number of rows: 1,000,000

Models trained or fine-tuned on lmsys/l...

tastypear/CausalLM-14B-DPO-al...
Text Generation • Upd... • 29.9k • 76

End of preview. Expand in Dataset Viewer.

< Previous 1 2 3 ... 10,000 Next >

Gated dataset You have been granted access to this dataset

YAML Metadata Warning: The task_categories "conversational" is not in the official list: text-classification, token-classification, table-question-answering, question-answering, zero-shot-classification, translation, summarization, feature-extraction, text-generation, text2text-generation, fill-mask, sentence-similarity, text-to-speech, text-to-audio, automatic-speech-recognition, audio-to-audio, audio-classification, voice-activity-detection, depth-estimation, image-classification, object-detection, image-segmentation, text-to-image, image-to-text, image-to-image, image-to-video, unconditional-image-generation, video-classification, reinforcement-learning, robotics, tabular-classification, tabular-regression, tabular-to-text, table-to-text, multiple-choice, text-retrieval, time-series-forecasting, text-to-video, image-text-to-text, visual-question-answering, document-question-answering, zero-shot-image-classification, graph-ml, mask-generation, zero-shot-object-detection, text-to-3d, image-to-3d, image-feature-extraction, other

LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset

This dataset contains one million real-world conversations with 25 state-of-the-art LLMs. It is collected from 210K unique IP addresses in the wild on the Vicuna demo and Chatbot Arena website from April to August 2023. Each sample includes a conversation ID, model name, conversation text in OpenAI API JSON format, detected language tag, and OpenAI moderation API tag.

tokyotech-llm/Llama-3.1-Swall…
Text Generation • Upda… • 18k • 15

tokyotech-llm/Llama-3.1-Swall…
Text Generation • Upd… • 15.6k • 16

tokyotech-llm/Llama-3.1-Swall…
Text Generation • Upd… • 8.94k • 14

jondurbin/bagel-8b-v1.0
Text Generation • Upd… • 8.04k • 52

jondurbin/bagel-34b-v0.2
Text Generation • Upda… • 6.5k • 40

Browse 367 models trained on this dataset

User consent is obtained through the "Terms of use" section on the data collection website. To ensure the safe release of data, we have made our best efforts to remove all conversations that contain personally identifiable information (PII). In addition, we have included the OpenAI moderation API output for each message. However, we have chosen to keep unsafe conversations so that researchers can study the safety-related questions associated with LLM usage in real-world scenarios as well as the OpenAI moderation process. We did not run decontamination on this dataset, so it may contain test questions from popular benchmarks.

For more details, please refer to the paper: https://arxiv.org/abs/2309.11998

Basic Statistics

Key	Value
# Conversations	1,000,000
# Models	25
# Users	210,479
# Languages	154
Avg. # Turns per Sample	2.0
Avg. # Tokens per Prompt	69.5
Avg. # Tokens per Response	214.5

PII Redaction

We partnered with the OpaquePrompts team to redact person names in this dataset to protect user privacy. Names like "Mary" and "James" in a conversation will appear as "NAME_1" and "NAME_2". For example:

```
Raw:      [ { "content": "Write me a bio. My Name is Mary I am a student
Redacted: [ { "content": "Write me a bio. My Name is NAME_1 I am a studen
```

Each conversation includes a "redacted" field to indicate if it has been redacted. This process may impact data quality and occasionally lead to incorrect redactions. We are working on improving the redaction quality and will release improved versions in the future. If you want to access the raw conversation data, please fill out the form with details about your intended use cases.

Uniqueness and Potential Usage

This dataset features large-scale real-world conversations with LLMs.

We believe it will help the AI research community answer important questions around topics like:

- Characteristics and distributions of real-world user prompts
- AI safety and content moderation
- Training instruction-following models
- Improving and evaluating LLM evaluation methods

- Model selection and request dispatching algorithms

For more details, please refer to the paper: https://arxiv.org/abs/2309.11998

LMSYS-Chat-1M Dataset License Agreement

This Agreement contains the terms and conditions that govern your access and use of the LMSYS-Chat-1M Dataset (as defined above). You may not use the LMSYS-Chat-1M Dataset if you do not accept this Agreement. By clicking to accept, accessing the LMSYS-Chat-1M Dataset, or both, you hereby agree to the terms of the Agreement. If you are agreeing to be bound by the Agreement on behalf of your employer or another entity, you represent and warrant that you have full legal authority to bind your employer or such entity to this Agreement. If you do not have the requisite authority, you may not accept the Agreement or access the LMSYS-Chat-1M Dataset on behalf of your employer or another entity.

- Safety and Moderation: **This dataset contains unsafe conversations that may be perceived as offensive or unsettling.** User should apply appropriate filters and safety measures before utilizing this dataset for training dialogue agents.
- Non-Endorsement: The views and opinions depicted in this dataset **do not reflect** the perspectives of the researchers or affiliated institutions engaged in the data collection process.
- Legal Compliance: You are mandated to use it in adherence with all pertinent laws and regulations.
- Model Specific Terms: When leveraging direct outputs of a specific model, users must adhere to its corresponding terms of use.

- Non-Identification: You **must not** attempt to identify the identities of individuals or infer any sensitive personal data encompassed in this dataset.
- Prohibited Transfers: You should not distribute, copy, disclose, assign, sublicense, embed, host, or otherwise transfer the dataset to any third party.
- Right to Request Deletion: At any time, we may require you to delete all copies of the conversation dataset (in whole or in part) in your possession and control. You will promptly comply with any and all such requests. Upon our request, you shall provide us with written confirmation of your compliance with such requirement.
- Termination: We may, at any time, for any reason or for no reason, terminate this Agreement, effective immediately upon notice to you. Upon termination, the license granted to you hereunder will immediately terminate, and you will immediately stop using the LMSYS-Chat-1M Dataset and destroy all copies of the LMSYS-Chat-1M Dataset and related materials in your possession or control.
- Limitation of Liability: IN NO EVENT WILL WE BE LIABLE FOR ANY CONSEQUENTIAL, INCIDENTAL, EXEMPLARY, PUNITIVE, SPECIAL, OR INDIRECT DAMAGES (INCLUDING DAMAGES FOR LOSS OF PROFITS, BUSINESS INTERRUPTION, OR LOSS OF INFORMATION) ARISING OUT OF OR RELATING TO THIS AGREEMENT OR ITS SUBJECT MATTER, EVEN IF WE HAVE BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Subject to your compliance with the terms and conditions of this Agreement, we grant to you, a limited, non-exclusive, non-transferable, non-sublicensable license to use the LMSYS-Chat-1M Dataset, including the conversation data and annotations, to research, develop, and improve software, algorithms, machine learning models, techniques, and technologies for both research and commercial purposes.

Citation

```
@misc{zheng2023lmsyschat1m,
      title={LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dat
      author={Lianmin Zheng and Wei-Lin Chiang and Ying Sheng and Tianle
      year={2023},
      eprint={2309.11998},
      archivePrefix={arXiv},
      primaryClass={cs.CL}
```