BEYOND THE DATA: ANALYSIS, FEATURE ENGINEERING AND BROWSER PLUGIN EXPANSION FOR THE SHARELM DATASET

Samer Attrah *
Google
Mountain view, CA, USA
samiratra95@gmail.com

September 28, 2025

ABSTRACT

As part of the Eleuther AI open AI summer research this year, we worked on expanding the ShareLM dataset browser extension, by adding support to multiple models in addition to redesigning some of the visual parts of the extension, in the mean time conducted several analysis and feature engineering on the ShareLM dataset to extract insight regarding the models, users, the conversations and the relations connecting them.

***K*eywords** Data analysis · Dataset · Natural language processing · feature engineering

1 Introduction

Instruction tuning is a method to improve and structure the information of pre-training a LLM, and the data used for this purpose needs to be high quality. Types of conversation that might take place are between **1)** human-human, **2)** AI model - AI model, and **3)** human - AI model, where the first is the most common type and can be found from conversation on the internet, messages, emails and social media besides many other platforms and formats, and usually get used in pre-training the models, while the AI model- AI model conversation might not be fit for use in fine tuning a model, since it might have all types of weaknesses a model can have from the lack of reason, keeping the topic, hallucinating, and other issues that might rise and not get noticed.

The third types is the conversations taking place between the AI model and user, on many occasions, most commonly in the commercial models websites such as ChatGPT [2], Claude [3], and HuggingFace chat [4]. on each of these websites there aare several model available for chatting, reasoning, searching the web, summarizing and analyzing long texts, and these conversations has a high quality due to the fact that they are managed by a human from one end, which help keep the conversation on the same topic and in a single context.

1.1 ShareLM chrome extension

Prior to this work a Google Chrome browser extension was built [1], to collect conversations from a group of websites and the LLMs available on them, including the ChatGPT, Claude and HuggingFaceChat model and any HuggingFace model integrated to a Gradio interface, where these conversations get detected from the browser interface element and recorded from the CSS selectors for the user message and the model message, and ordered in a format of a prompt-response conversations for as many turns as the user uses the LLM.

[*] Not affiliated at the time of publishing the research

[2] https://chatgpt.com/

[3] https://claude.ai/new

[4] https://huggingface.co/chat

Feature	Description
Conversation ID	Autonomously generated encrypted phrase, unique for every conversation.
Conversation	The contents of the textual human-model conversation.
Model name	The display name of the model (ChatGPT, Claude, Gemini).
User ID	Autonomously generated encrypted phrase, unique for every user.
Time stamp	The time when the conversation took place
Source	the source of the conversation, and from which dataset it was taken
User metadata	further information and details about the user
Conversation metadata	further details and description for the conversation

Table 1: ShareLM dataset features

In addition to the conversation some metadata get collected about the user, such as the age, and gender, and for the conversation the rate, language and toxicity. besides other features, that part of it needs to be provided manually by the user and others get collected autonomously by the browser extension.

Another main part of the browser extension is the user interface, and banner which indicates the operability of the extension and enable the user to fill some of the metadata in addition to other uses.

1.2 ShareLM dataset

The ShareLM dataset [5] is available in it's latest version on HuggingFace with around 350K human - model conversations, partially were collected using the browser extension and another part were added from datasets such as Anthropic/hh-rlhf and PRISM-alignment, besides many other sources and datasets.

The dataset conversations are in a wide range of languages, turn counts and lengths. and they have a wide variety of topics and arguments. and the features of the dataset are as in the Table 1

The two metadata features, have many sub-features can be found listed in Tables 2 and 3

Sub-feature	Description
Location	Geographical location of the user
Age	An integer for representing the users age
Gender	The gender of the user

Table 2: User metadata

Sub-feature	Description
Rate	Zero(s) or one(s) numerically representing a like or a dislike, respectively, assigned by the user to the conversation.
Language	The language used in the conversation.
Redacted	Flag indicating if any content in the conversation was redacted for privacy/safety.
Toxic	The value measured for the toxicity of the conversation.
Title	The title of the conversation.
Custom instruction	a boolean to indicate if a custom instruction (or system message) was used when the conversation was made.
Status	indicate the conversation's progress through quality control pipeline after being collected.

Table 3: Conversation metadata

The main contribution of this work is that we improve many aspects of the ShareLM extension and dataset:

1. Add support to several models, and websites for the extension to collect user - model conversations from.
2. Improve the banner design to be more efficient and fit a wider range of websites.

[5] `https://huggingface.co/datasets/shachardon/ShareLM`

3. Add a new synthetic sub-feature to the conversation metadata, to classify the topic of the conversation, in an automated process.
4. Analyze the dataset to find best way possible to improve the browser extension and the dataset, with possibility for driving social and business insights.

The next parts of the article are structured in the sections: 2 Analysis and expansion, 4 Possible automations, 5 Conclusion.

2 Analysis and expansion

When collecting a dataset, having a larger number of features and a wider range of representation for the real-life experiences and use cases, enriches the distribution of the dataset and enable the model that trains on it to have better performance on more tasks and problems, and based on that several improvements were made.

2.1 Chrome extension expansion

An expansion to the chrome extension, using Jules: an async development agent by Google [6], included adding more websites and models for the list of source of conversations collected, and fit a wider range of users and preferences.

The models were added are: Gemini [7], MistralAI chat [8], Perplexity [9], Poe [10] and Cohere playground chat [11],

In addition to the platform that includes many models, especially Poe, minor changes to the extension banner design and GitHub [12] repository was applied.

2.2 Dataset analysis

In the dataset the human *user* does not get identified by their name or any other personal information but by a User ID generated by an algorithm that will produce the same ID every time provided the same name, while the other data included for the user which get manually and optionally inputted by the user, are *location*, *age*, and *gender*, and using this data can find the activity of the extension and models according to the location, age group and gender, but due to the fact that it is not a required information from the user, only few insertions to the dataset includes this information, which make make any conclusions based on them invalid.

For the information included about the *model*, only the Model name is available, and it can be used to derive many insights when combined with the information about the users and the conversations, but that is not possible due to the inaccuracy in registering the model name from the website such as inserting "LMArena.ai" for all the models available on the platform without specifying the type of the model in use.

The Third main element of the dataset is the conversation, which makes the majority of the data, where the provided information is a conversation ID, as mentioned in the Table 1, and the conversation is fully recorded as prompt-response pairs between the user and the model, which can be regarding any topic or in any type of language, and in any length,

In addition to the conversation metadata mentioned in Table 3 we autonomously added a *synthetic* eighth sub-feature to the conversation metadata to describe the "**topic**" of the conversation, and that came because having a better understanding for the conversation content will enable further improvements, especially when observed against other features, and adding this feature included around 10K rows of the dataset, for this task five classes were selected [2] to describe the topic:

- Assisting/creative writing.
- Analysis/decision.
- Explanation/coding.
- Factual info.

[6] https://jules.google/

[7] https://gemini.google.com/app

[8] https://chat.mistral.ai/chat

[9] https://www.perplexity.ai/

[10] https://poe.com/

[11] https://dashboard.cohere.com/playground/chat

[12] https://github.com/shachardon/share-lm

- Math reason

Classifying the conversations and labeling them with one of these classes was done by a **gemma-3n-e2b-it**, prompted using the Google GenAI library, and the full code can be found in the data analysis repository [13], choosing this model because it is the best and latest in its size and type, which enable it to classify, 15K conversations everyday, with minimal resources requirement.

2.3 Visualizations

Using this improved dataset, and after extracting the rows that includes a non-empty value for the *model name* which sums up to around 10K, and after adding the topic sub-feature to them, created a group of visualizations as follows:

2.3.1 Contributions per user

the number of contribution made by each user, by counting the conversations submitted by each User ID, which can show that only nine users has a significant number of contribution with the highest reaching less than 900, and as mentioned earlier we can not find who the person is since the user name is not provided but the User ID is what included in the dataset, and as in figure 1

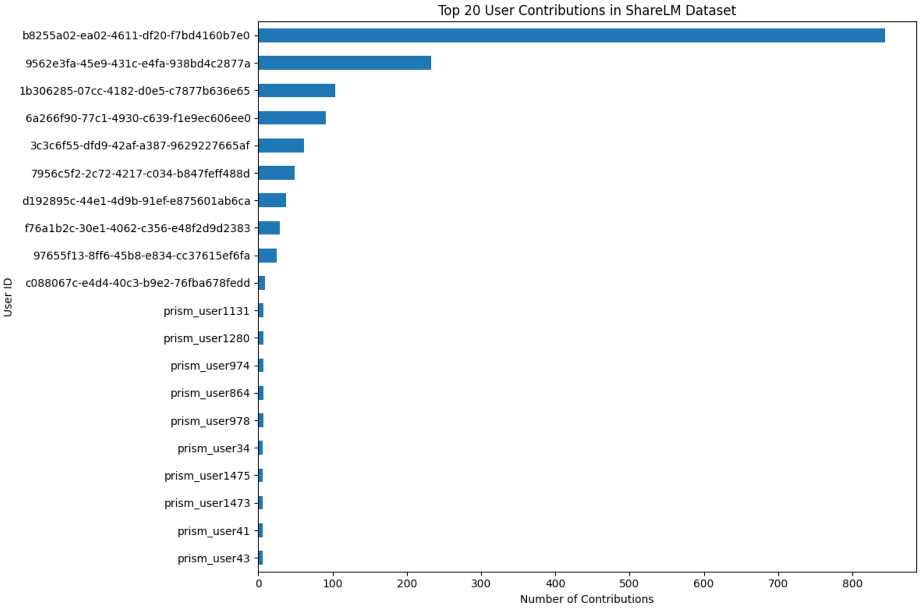

Figure 1: Contribution per user

The plot indicate that the retention of the users is so limited today, and only a few of them are significantly contributing, which can also result in having a limited number of topics and smaller number of models with higher frequency and usage, which indicates the requirement for a more attractive activity, for example redesigning the extension, promoting it, and adding incentives can make the users more interested in contributing and interacting with it.

2.3.2 Conversation length

one measure that is most important for model training engineers is the knowing of the lengths of the conversations included in the dataset, which get measured by the *number of turns* i.e. pairs of prompt-response between the user and the model, and there are other methods of measurements such as the number of characters, number of tokens, or the number of words.

The plot 2 for the conversation length analysis, can indicate the following facts about the data sample analyzed:

- The longest conversation is a single one with around a hundred turns and the shortest has a turn only.

[13] `https://github.com/Samir-atra/share-lm_dataset_analysis`

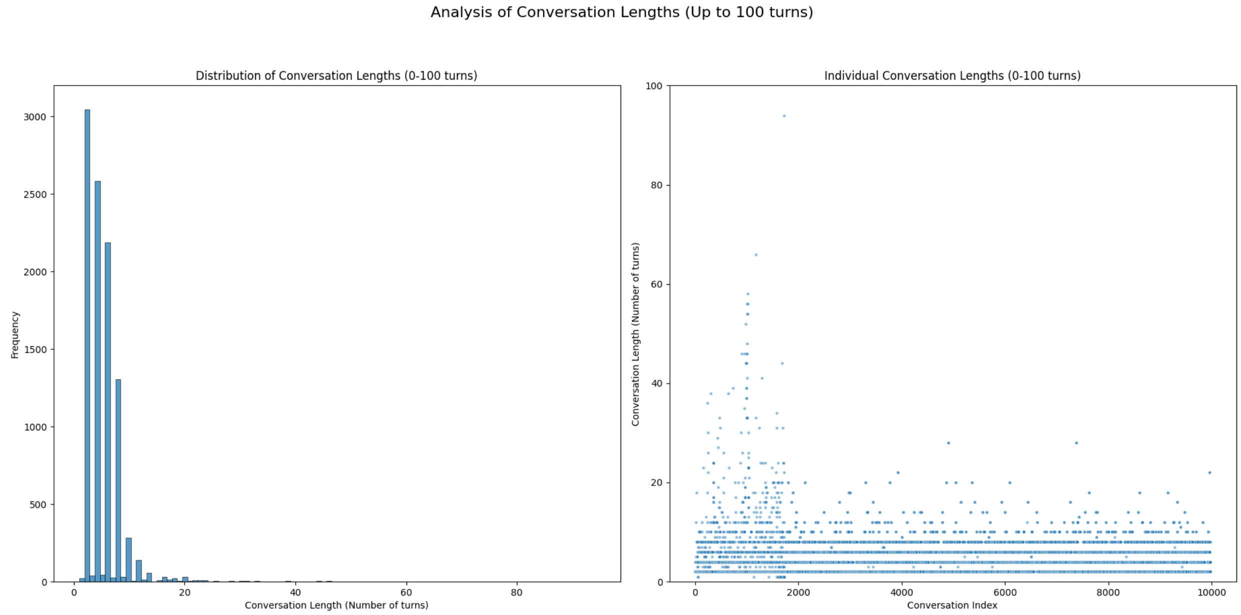

Figure 2: Conversation length, **Left**: the conversation number of turns against the frequency, **Right**: a scatter plot for the conversation number of turns against the conversation index

- The majority of the conversations are below 20 turns.
- The turn count with highest occurrences is the 2 turns conversation reaching around 3000 in frequency.

these findings can indicate the high efficiency for the models used and the users are of a specific group instead of a wide range.

2.3.3 Model counts

The model counts can help in understanding the users of the plugin and which model they prefer, and currently the model names are not accurate and as mentioned in a previous section.

The plot in figure 3 shows:

- One model which is GPT4 is dominating the use with around 250 conversations.
- The more frequently used models are only 12 and 6 of them are GPT models by OpenAI.
- LMArena.ai [14] is heavily used being in the second place, although tens of models on the website all of them are under the same model name.

2.3.4 Topic counts

For the topic counting and visualization the classes were used are the same as in the "WildChat" dataset and they are as specified above, applying the classification on the conversations of the dataset can show the following:

- Analysis/decision and factual info classes are the most common in the dataset with more than 7000 conversations in these classes.
- Assistance and deployment classes are not part of the criteria but can be found in one occurrence each in the dataset.
- Math reason is almost not existing in the dataset.
- "explanation" and "creative writing" have separate frequencies and get summed with the main classes, to have more accurate and representative numbers.

[14] https://lmarena.ai/

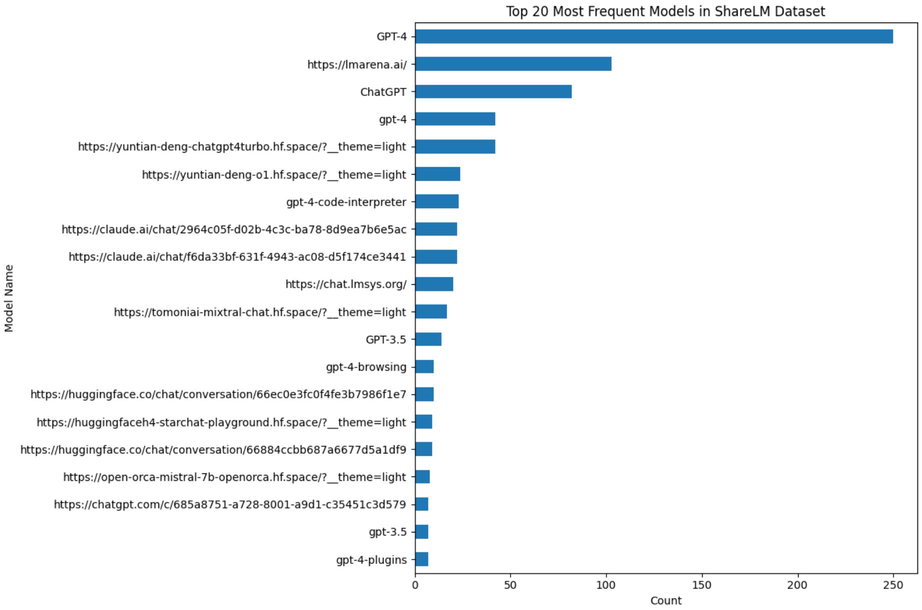

Figure 3: Model use frequency

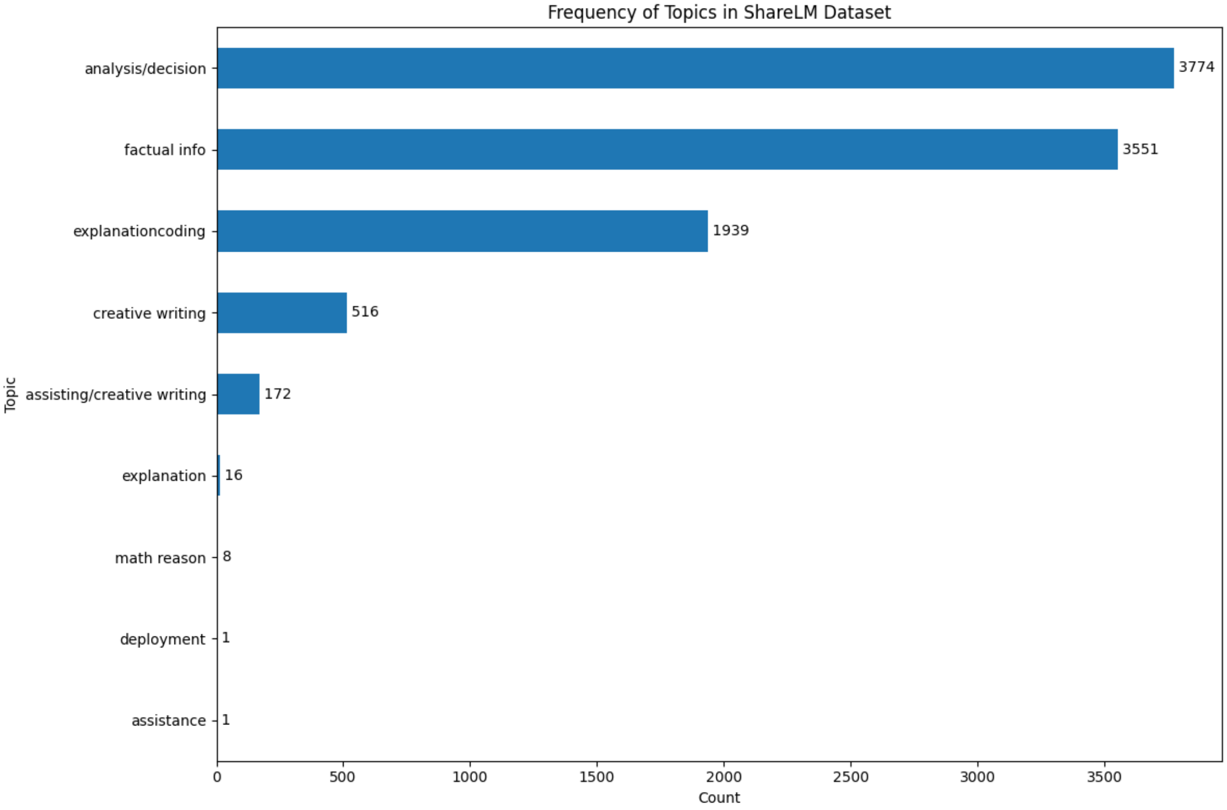

Figure 4: topics frequency

2.3.5 Topic per model

To better understand the user patterns and how each model is related to a conversation topic or a use case, conducted an analysis of the topic vs. model name, as in figure 5

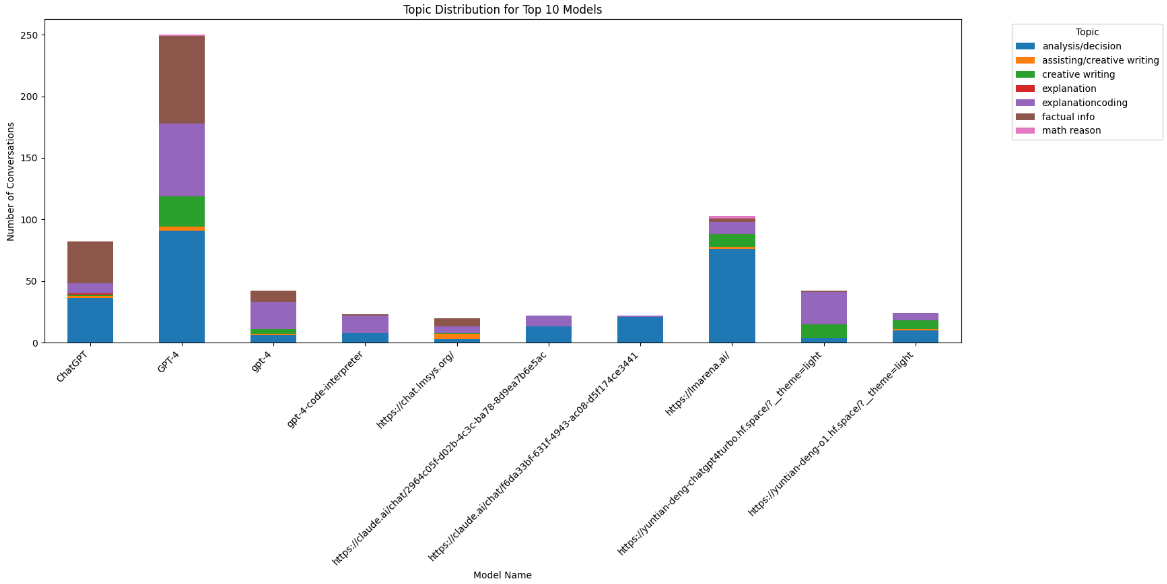

Figure 5: Topics for each model

- GPT4 which has the highest number in usage frequency is used for finding factual info, analysis/decision and explanation/coding.
- LMArena is mostly used for analysis/decision, which has the second highest usage frequency.
- ChatGPT comes third with a topic pattern similar to the GPT4 model which indicates that a similar user type are the ones interested in GPT models.

These insights shows that the main use case of the model is analysis and decision making assistance, which gives a general category for the dataset, and reflects the interest of users, in addition to providing understanding and drive future development for the extension.

3 Suggested improvements

The structure of the dataset is fully comprehensive and have sufficient number of features and sub-features to be useful for instruction tuning a model, for the widest range of applications, in addition to being a resource of information for studies and research in humanities.

Starting with the browser extension and to improve the efficiency of the user metadata collection have a fully annotated dataset and more useful set of features, the upgrade of the design of the browser extension to have the user metadata insertion as a requirement instead of having it as an optional addition.

Due to the importance of the model name feature as part of the dataset for model training and for further research, and the increasing number of models available, we suggest changing the detection of the model name to be from the user interface temporarily and have an accurate model name feature, instead of collecting the name from the backend of the website and continue looking for a better approach.

Since the speed of collecting the data and the size of the dataset is a major challenge, improving the User interface design to be more attractive and simpler for the users, can be a crucial goal to drive further advancement.

To have a wider range of features and increase engagement from a broader user base, adding support to models on educational platforms and AI tutors should be of interest, in addition to collecting conversations from help and website navigation assistants, integrated to all types of websites.

4 Possible automation

Annotating the data to have a complete and understandable set of instructions to be used for training, is an essential part of creating the dataset, but due to the long time and amount of effort required, we suggest automating the process by using fine-tuned LLMs tested for high performance in the application, and classification, which will deliver less accurate but quick results, clarifying the path for more structured improvement to the dataset and browser extension. or could be used in studies and estimations that do not necessitate high accuracy.

4.1 Adding features

A few features of the dataset, that are empty today can be completed by having access and understanding to the conversations in the dataset, which can be obtained with a large language model, possibly a commercial model or a one fine tuned for this purpose.

4.1.1 Language classification

The "language" sub-feature that can be found in Table 3 does not have content today, and requires to be completed, and since multilinguality is a well developed large language model capability then iterating on the conversations, with a prompt for inferencing the language, can help improving the usability of the dataset.

4.1.2 Creating titles

Similar to the process described in the previous sub section, completing the "title" sub-feature can give a shortcut to understanding the conversations, and serving as an extra short summary, which might be used for fine tuning the models on another task, that is popular now a days.

4.1.3 Completing the model name

Instead of inferencing the language and creating a title based on the conversation, can use the "source" sub-feature to infer the "model name" sub feature, especially knowing that many datasets were added to the ShareLM dataset, generated by known models mentioned in their research articles.

4.2 Extracting further insights

Further insights can be extracted from the dataset in case of having a complete metadata and all the features, which can be used in developing a deeper understanding to the users and drive the future improvements of the browser extension and the models used:

- Related to the **users**: finding each users favorite model and how often they use it or switch to use another, and looking for patterns of use for each user regarding the time of usage. In addition to finding the location of the widest group of users, the most contributing age group and gender.
- Regarding the **model**: Find the rate-to-model ratio (user engagement), which regardless of the rating value the can be used to assess the model, can improve the understanding of the user behavior of each model and how interactive they are. besides that having an accurate model name can help estimate the toxicity-per-model, and which model needs improvement in this regard. and finding the language-to-model use while having a rating by the user can help evaluate the multilingual performance of the models, supported by the regional favorite and the age group using the specific model.
- For the **conversation**: visualizing the rate against the language can give an indication for the multilinguality performance for a certain model, while visualizing the language against the toxicity, might also help in measuring the toxicity for a specific model in a language for a study in machine translation and multilinguality, in addition to other studies that can be found by visualizing one of the conversation metadata against one of the feature of the dataset and according to the use case it is done for.

5 Conclusion

The research presents a path for collecting and creating datasets that can be used for post-training models, in addition to the possibility of using the dataset in studying and research in humanities and social scientific fields. It is more or less a methodology for understanding the commercial LLM world, in addition to being a model training material. inspired by google analytics and google trends, with a more general and raw content collected from a point closer to the user, or sometimes manually from them. in addition to suggesting a group of automation methods for expanding web applications and datasets.

References

[1] Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. The sharelm collection and plugin: contributing human-model chats for the benefit of the community. *arXiv preprint arXiv:2408.08291*, 2024.

[2] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*, 2024.