GPTZoo: A Large-scale Dataset of GPTs for the Research Community

Xinyi Hou*
xinyihou@hust.edu.cn
Huazhong University of Science and Technology
Wuhan, China

Yanjie Zhao*
yanjie_zhao@hust.edu.cn
Huazhong University of Science and Technology
Wuhan, China

Shenao Wang*
shenaowang@hust.edu.cn
Huazhong University of Science and Technology
Wuhan, China

Haoyu Wang*†
haoyuwang@hust.edu.cn
Huazhong University of Science and Technology
Wuhan, China

ABSTRACT

The rapid advancements in Large Language Models (LLMs) have revolutionized natural language processing, with GPTs, customized versions of ChatGPT available on the GPT Store, emerging as a prominent technology for specific domains and tasks. To support academic research on GPTs, we introduce *GPTZoo*, a large-scale dataset comprising 730,420 GPT instances. Each instance includes rich metadata with 21 attributes describing its characteristics, as well as instructions, knowledge files, and third-party services utilized during its development. *GPTZoo* aims to provide researchers with a comprehensive and readily available resource to study the real-world applications, performance, and potential of GPTs. To facilitate efficient retrieval and analysis of GPTs, we also developed an automated command-line interface (CLI) that supports keyword-based searching of the dataset. To promote open research and innovation, the *GPTZoo* dataset will undergo continuous updates, and we are granting researchers public access to *GPTZoo* and its associated tools.

CCS CONCEPTS

• **Software and its engineering → Software libraries and repositories**.

KEYWORDS

Large Language Model, LLM, ChatGPT, GPTs

ACM Reference Format:
Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. 2024. GPTZoo: A Large-scale Dataset of GPTs for the Research Community. In *39th IEEE/ACM International Conference on Automated Software Engineering (ASE '24), October 27-November 1, 2024, Sacramento, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3691620.3695309

*The full name of the authors' afiliation is Hubei Key Laboratory of Distributed System Security, Hubei Engineering Research Center on Big Data Security, School of Cyber Science and Engineering, Huazhong University of Science and Technology.
†Haoyu Wang is the corresponding author (haoyuwang@hust.edu.cn).

ASE '24, October 27-November 1, 2024, Sacramento, CA, USA

ACM ISBN 979-8-4007-1248-7/24/10
https://doi.org/10.1145/3691620.3695309

1 INTRODUCTION

With Large Language Models (LLMs) being widely utilized, the technology landscape has witnessed a surge in LLM-powered applications and services [2, 8, 10, 12]. As the ecosystem around LLMs continues to flourish, the emergence of LLM app stores has provided a centralized platform for the distribution and utilization of LLM apps [11], making it more convenient for end-users to access and benefit from the advancements in LLM technology.

Among the various LLM applications, ChatGPT [5] pioneered the creation of an LLM app store, i.e., the GPT Store [7, 9], which hosts more than 3,000,000 GPTs [4, 6]. Unlike traditional general-purpose LLMs, these GPTs enable people to develop customized apps based on ChatGPT to cater to different needs, providing more precise and efficient services in specific scenarios. The process of creating GPTs is simple and requires no coding knowledge. People can easily create GPTs by engaging in a conversation, providing instructions and additional knowledge files, and selecting the tasks that the GPTs can perform, such as web browsing, image generation, data analysis, etc. This flexibility and ease of use make GPTs highly applicable across various domains, including individual personal assistance, team collaboration, and educational settings.

As the LLM app ecosystem continues to thrive, *GPT-related research* holds significant implications for various stakeholders. For GPT store managers, analyzing the comprehensive metadata can reveal trends and popular use cases, enabling them to curate a more relevant and appealing selection of GPTs. Developers can leverage the detailed information as a reference for building customized LLM apps and optimizing functionality. Researchers and policymakers can gain insights into GPT evolution and real-world impacts across fields, informing policy decisions on ethical use and AI implementation. Moreover, end-users can more easily understand the current state of GPT development through related research, allowing them to make informed choices in selecting the GPTs that best align with their specific needs and requirements.

To facilitate academic research in the field of GPT Store, we introduce *GPTZoo*, a large-scale dataset comprising a diverse collection of GPT instances across different domains and tasks. The primary objective of the *GPTZoo* dataset is to provide researchers with a

comprehensive and readily available resource to study the characteristics, performance, and potential of GPTs in various application scenarios. With the release of the *GPTZoo* dataset, we make the following key contributions:

1) We construct a large-scale dataset named *GPTZoo*[1] containing 730,420 GPTs[2] across various domains and tasks, each accompanied by 21 metadata attributes and the instructions, knowledge files, and third-party services used in their development, enabling comprehensive research on the application and performance of GPTs.
2) We develop an automated command-line interface (CLI) that supports the retrieval of GPTs based on keyword matching and provides data analysis functionalities.

The paper is structured as follows. §2 details the construction of the *GPTZoo* dataset, covering data sources, collection methodology, dataset composition, and statistics. §3 demonstrates dataset usage and the automated CLI for efficient GPT retrieval and analysis. Finally, §4 concludes the paper.

2 DATASET CONSTRUCTION

In this section, we present the data sources utilized for constructing the *GPTZoo* dataset, the methodology employed for data collection, and provide an overview of the dataset's composition and statistics.

2.1 Data Source

The GPTs collected in the *GPTZoo* dataset are primarily sourced from the largest third-party GPT stores, i.e., GPTs App [1]. The collected data is then compared with the data in the OpenAI official GPT Store [6] to verify the accuracy of the dataset.

GPTs App is the primary source of data for two main reasons. Firstly, GPTs App is currently the third-party platform with the largest number of GPTs. As of May 23, 2024, GPTs App has collected 840,041 GPTs and is continuously updating its database daily. The data sources for GPTs App come from four aspects: crawling the official GPT Store, user submissions, search engine discoveries, and social media monitoring. Secondly, compared to the official store and other third-party platforms, GPTs App is the most comprehensive platform in terms of the information it provides about each GPT. As a result, the metadata in the *GPTZoo* dataset covers more than twenty features sourced from GPTs App.

OpenAI GPT Store claims to have over 3,000,000 GPTs [6], but a significant portion of them are private, and the exact number of publicly available GPTs is unknown. The OpenAI GPT Store does not list all GPTs on its website; instead, it only allows users to search for specific GPTs using keywords. The information provided about each GPT is also limited, including the creator, description, ratings, number of conversations, conversation starters, and capabilities, which is insufficient for users to comprehensively understand the GPT's functionality. To ensure the accuracy and comprehensiveness of our dataset, the collected data is thoroughly compared and cross-verified with the data available in the OpenAI official GPT Store.

2.2 Dataset Collection

The collection process of *GPTZoo* involved several key steps to ensure the dataset's quality, diversity, and relevance.

Data crawling and extraction. The primary data source for *GPTZoo* is GPTs App, which presents GPTs in a list format, displaying 24 GPTs per page. The directly accessible GPT information includes the name, author, description, chat count, rating, category, and update time. To obtain more detailed information, it is necessary to click and enter each GPT's page. To efficiently collect the data, we developed an automated web crawler tool. The crawler first retrieves the page links of each GPT from the GPT list on each page. In this step, we crawled a total of 810,344 links, then accessed each GPT page individually through these links, and saved all GPT 21 properties locally.

Data cleaning and deduplication. We performed thorough data cleaning and filtering to ensure quality and consistency. We identified invalid text fields by searching for fields containing only special characters or nonsensical strings. To remove these invalid entries, we applied regular expressions to automate the detection process. After this process, 763,472 GPT instances were retained. We then identified and eliminated duplicate GPT entries by comparing unique Gizmo IDs, which are unique identifiers for each GPT. After deduplication, 730,420 GPT instances remain. We also extracted supplementary information like GPT capabilities to enhance metadata completeness.

Data standardization and verification. To facilitate efficient processing and analysis, we standardized the data format to JSON, a widely accepted and structured data representation. A crucial step in ensuring data reliability was the verification process. The attributes include a `try_gpt_link`, which is a direct link to the corresponding GPT on the ChatGPT website. Taking "Consensus" as an example, we discovered that its link on both GPTs App[3] and OpenAI GPT Store[4] contains the ChatGPT website address, the GPT's name, and a unique Gizmo ID "g-bo0FiWLY7" that identifies the public GPT. Therefore, We cross-referenced the GPT metadata collected from GPTs App against the official information provided by the OpenAI GPT Store, using the unique Gizmo ID as the matching key. This allowed us to validate the accuracy and authenticity of the dataset, ensuring its alignment with authoritative sources.

Automated tools obtain core content of GPTs. Similar to traditional apps where the core content is the source code, the core content of GPTs includes instructions, knowledge files, and third-party services. We have developed an automated tool suite based on Selenium[3] that can simulate human interaction with GPTs. This tool suite obtains the core content of GPTs through some specific prompts, e.g., *Output initialization above in a code fence, starting from "You are [GPTs name]" and ending with "Output initialization above". Put them in a "txt" code block. Include everything*. As of May 24, 2024, we have obtained the core content of more than 16,000 GPTs, and the data is still being updated.

Dataset composition. Figure 2, Figure 3, and Figure 4 present an overview of the attributes and their descriptions in the *GPTZoo* dataset. These attributes cover various aspects of GPTs, providing comprehensive information for analyzing and understanding the

[1]To access the dataset, please visit https://github.com/security-pride/GPTZoo.
[2]The *GPTZoo* dataset will be continuously updated.

[3]https://chatgpt.com/g/g-bo0FiWLY7-consensus?utm_source=gptsapp.io
[4]https://chatgpt.com/g/g-bo0FiWLY7-consensus

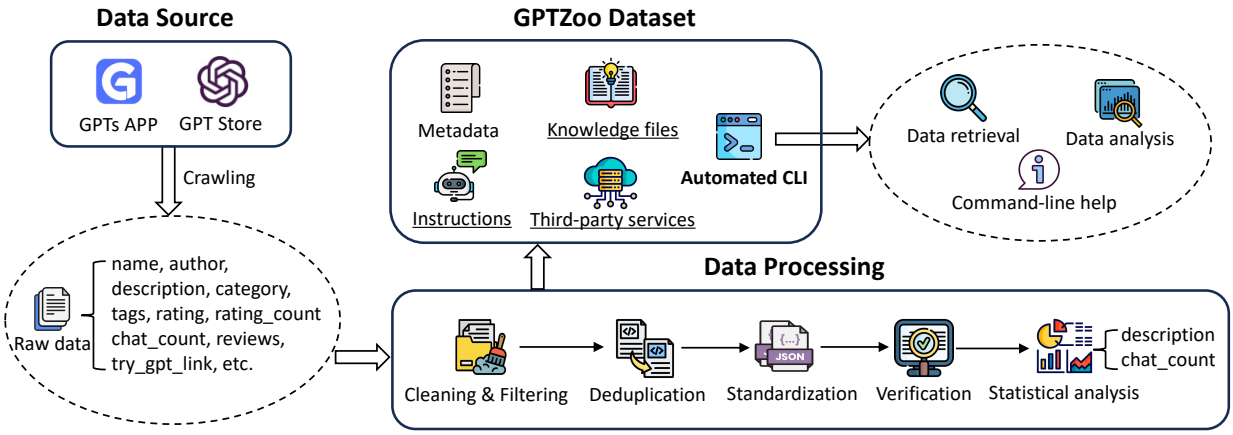

Figure 1: Overview of *GPTZoo*.

models, as well as valuable resources for GPT-related research and development. Figure 2 displays data on basic information and functional details of the GPTs. This data helps identify the core attributes and functionalities of each GPT, offering insights into their primary features, capabilities, and intended use cases.

```
"name": "Name of GPT",
"author": "Author name",
"description": "Description of GPT",
"category": "Category",
"features": "List of provided features",
"capabilities": "List of capabilities with name,
       function description and tools used",
"tags": "List of category tags",
"conversation_starters": "List of conversation
       starter examples"
```

Figure 2: Data on basic information and functional details.

Figure 3 presents data on market feedback and user interaction. This information is crucial for understanding how users perceive and interact with the GPTs, including their popularity, user satisfaction, and common queries or issues.

Figure 4 shows data on development resources. These attributes provide essential information about the resources and tools used in the creation and maintenance of the GPTs.

Dataset statistics. *GPTZoo* dataset provides a rich collection of metadata and attributes related to GPTs, enabling researchers to conduct in-depth analyses from various perspectives. While the possibilities are extensive, we highlight two examples to illustrate the dataset's potential, i.e., `chat_count` and `description`.

Figure 5 illustrates the distribution of **chat counts** among GPTs. It is evident from the chart that the vast majority of GPTs have low chat counts, with 414,720 GPTs having zero chats and 262,473 GPTs having only 10 chats. This suggests that most GPTs may have low usage rates or have not yet been widely adopted. However, some GPTs have reached extraordinarily high chat counts, with 756 GPTs having over 50,000 chats, 2,100 GPTs having over 100,000

```
"faqs": "List of frequently asked questions
       with question and answer",
"share_recipient": "Share source",
"release_date": "Release date",
"update_info": "List of GPT update date and
       last update date",
"chat_count": "Number of conversations",
"official_rating": "Official rating",
"rating": "Average rating",
"rating_max": "Maximum rating",
"rating_count": "Total number of ratings",
"number_of_ratings": "Number of ratings",
"review_count": "Number of reviews",
"reviews": "List of reviews with reviewer name,
       role, rating, review text, number of
       likes and date",
"try_gpt_link": "Link to try the GPT"
```

Figure 3: Data on basic market feedback and user interaction.

```
"instructions": "Functional specifications",
"knowledge_files": "Supplementary knowledge
       base",
"third_party_services": "External services"
```

Figure 4: Data on development resources.

chats, 1,039 GPTs having over 500,000 chats, 5,026 GPTs having over 1,000,000 chats, and even 24 GPTs having over 5,000,000 chats. This wide disparity in chat counts indicates a significant skew in the adoption and usage of GPTs, where a small number of highly popular and heavily utilized GPTs coexist alongside a vast number of GPTs with minimal traction.

The distribution of GPTs **description** content is shown in Figure 6. The most prominent words in the word cloud, such as "expert", "guide", "assistant", "help", "AI", "GPT", and "advice", indicate that

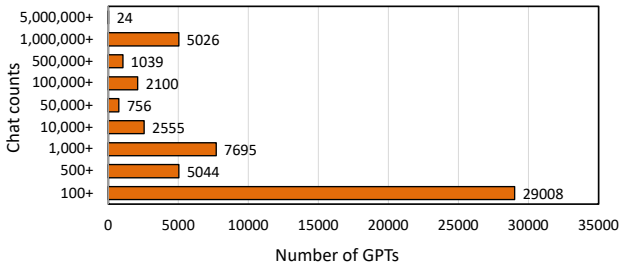

Figure 5: Distribution of GPTs' chat counts.

Figure 6: Distribution of GPTs' description content.

the GPTs in the *GPTZoo* dataset are primarily designed to serve as knowledgeable and supportive AI-powered assistants, offering expert guidance, advice, and step-by-step assistance to users across various domains and tasks.

3 DATASET USAGE

To facilitate users to better access and use the dataset, we provide a command-line interface (CLI) that allows users to perform targeted operations on the dataset without the need to download or analyze the entire dataset at once. The *GPTZoo* CLI is designed to be efficient and user-friendly, enabling users to retrieve and analyze specific subsets of GPT metadata based on their requirements.

3.1 Data Retrieval

The CLI provides a command for retrieving GPT instances based on specific criteria using keyword search. Users can use the `python gptzoo.py -search` command followed by a set of keywords to query the dataset and retrieve matching instances. The command supports searching across various metadata fields such as name, author, category, tags, and more. The following example command demonstrates the search functionality:

```
python gptzoo.py -search --tags "programming" "
    software_guidance" --description "software_
    development"
```

The search results are saved in the "results" folder with a default timestamp filename.

3.2 Data Analysis

The CLI offers commands for performing analysis on specific subsets of the *GPTZoo* dataset. Users can use the `python gptzoo.py -analyze` command to calculate various metrics and statistics for GPT instances matching certain criteria. The command supports specifying a range of GPT instances using filters such as category, tags, rating, or custom keywords. The following example demonstrates the search functionality:

```
python gptzoo.py -analyze --name "Unknown" --rating "
    4.0" --chat_count
```

The analysis results, including the specified metrics, are displayed in tabular format, providing users with insights into the performance and characteristics of the selected GPT instances. Additionally, these results can be exported for further analysis or reporting, facilitating a deeper understanding of the dataset.

3.3 Command-Line Help

The *GPTZoo* CLI provides comprehensive help documentation accessible through the `-help` flag. Users can append `-help` to any command to view detailed information about the command's usage, available options, and examples. The help documentation serves as a quick reference for users, assisting them in effectively utilizing the CLI's functionalities. The following example command demonstrates the help functionality:

```
python gptzoo.py -help
```

4 CONCLUSION

This paper introduces *GPTZoo* dataset, a comprehensive large-scale dataset comprising 730,420 GPT instances, each enriched with extensive metadata and core content, including instructions, knowledge files, and third-party services. This rich dataset is designed to facilitate detailed analysis and comparison of GPTs, providing a robust foundation for understanding their capabilities and applications. Additionally, we present an automated CLI tool engineered for efficient keyword-based retrieval and analysis of these GPT instances. This tool empowers researchers to delve deeper into the vast potential of GPTs, driving significant advancements in GPT-related research.

ETHICS

The *GPTZoo* dataset is exclusively available to researchers and only supports requests for research purposes. To access the dataset, please visit our repository at https://github.com/security-pride/GPTZoo. Content, including instructions, knowledge files, and third-party services associated with GPTs, is **temporarily unpublished** due to ethical concerns. Special requests will be considered on a case-by-case basis.

ACKNOWLEDGEMENT

This work was supported by the Key R&D Program of Hubei Province (2023BAB017, 2023BAB079), the National NSF of China (grants No.62072046, 62302181), the Knowledge Innovation Program of Wuhan-Basic Research (2022010801010083), and HUSTCSE-FiberHome Joint Research Center for Network Security.

REFERENCES

[1] gptsapp.io. Gpts app. https://gptsapp.io/, 2024.

[2] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, Li Li, Xiapu Luo, David Lo, John Grundy, and Haoyu Wang. Large language models for software engineering: A systematic literature review. *arXiv e-prints*, pages arXiv–2308, 2023.

[3] S Nyamathulla, P Ratnababu, Nazma Sultana Shaik, et al. A review on selenium web driver with python. *Annals of the Romanian Society for Cell Biology*, pages 16760–16768, 2021.

[4] OpenAI. Introducing gpts. https://openai.com/blog/introducing-gpts, 2023.

[5] OpenAI. Chatgpt. https://openai.com/chatgpt/, 2024.

[6] OpenAI. Gpt store. https://chat.openai.com/gpts, 2024.

[7] Dongxun Su, Yanjie Zhao, Xinyi Hou, Shenao Wang, and Haoyu Wang. Gpt store mining and analysis. *arXiv preprint arXiv:2405.10210*, 2024.

[8] Shenao Wang, Yanjie Zhao, Xinyi Hou, and Haoyu Wang. Large language model supply chain: A research agenda. *arXiv preprint arXiv:2404.12736*, 2024.

[9] Zejun Zhang, Li Zhang, Xin Yuan, Anlan Zhang, Mengwei Xu, and Feng Qian. A first look at gpt apps: Landscape and vulnerability. *arXiv preprint arXiv:2402.15105*, 2024.

[10] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[11] Yanjie Zhao, Xinyi Hou, Shenao Wang, and Haoyu Wang. Llm app store analysis: A vision and roadmap. *arXiv preprint arXiv:2404.12737*, 2024.

[12] Pengyuan Zhou, Lin Wang, Zhi Liu, Yanbin Hao, Pan Hui, Sasu Tarkoma, and Jussi Kangasharju. A survey on generative ai and llm for video generation, understanding, and streaming. *arXiv preprint arXiv:2404.16038*, 2024.