# Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias

**Yue Yu** [*1] **Yuchen Zhuang** [*1] **Jieyu Zhang** [*2] **Jiaming Shen** [3] **Yu Meng** [4] **Chao Zhang** [1]

## Abstract

Large language models (LLMs) have been recently leveraged as training data generators for text classification. While previous research has explored different approaches to training models using generated data, there is a tendency to rely on simple class-conditional prompts, which may limit the diversity of the generated data and inherit systematic biases of LLM. Thus, we investigate training data generation with diversely attributed prompts (e.g., specifying the length and style), which have the potential to yield diverse and attributed generated data. Our investigation focuses on datasets with high cardinality and diverse domains, wherein we demonstrate that attributed prompts outperform simple class-conditional prompts in terms of the resulting model's performance. Additionally, we present a comprehensive empirical study on data generation encompassing vital aspects like bias, diversity, and efficiency. Importantly, our findings highlight two key observations: firstly, synthetic datasets generated by simple prompts exhibit significant biases, such as regional bias; secondly, attribute diversity plays a pivotal role in enhancing model performance.

## 1 Introduction

Large language models (LLMs) have achieved remarkable performance on a variety of natural language processing (NLP) tasks (Devlin et al., 2019; Brown et al., 2020; Ouyang et al., 2022; Liang et al., 2022; OpenAI, 2023a;b). Among many applications of LLMs, researchers have recently purposed them as training data generator, especially for the long-standing task of text classification (Ye et al., 2022a; Gao et al., 2023; Ye et al., 2022b; Meng et al., 2022; Yu et al., 2023; Chen et al., 2023), in order to eliminate the need for task-specific data and annotations. While existing efforts have shown the applicability of LLM as data generator, they typically focus on advancing the training stage where the generated data are used to train a task-specific model, leaving the upstream data generation process underexplored. In particular, a simple class-conditional prompt is adopted to query LLM for data generation, which hinders the diversity of the generated data (Chen et al., 2023; Tevet & Berant, 2021) and possibly inherits the generative bias of LLM (Ferrara, 2023; Zhuo et al., 2023; Kirk et al., 2021).

In this work, we instead anchor on the data generation part and aim to generate data with diversely attributed prompts in order to remedy the issue of low diversity and inherited bias due to simple prompts. We ground the LLM to ChatGPT for its ability to generate high-quality, human like text, and consider four challenging topic classification tasks from various domains with high cardinality. For each dataset, we first identify data attributes as well as the corresponding attribute values with the assistance of the ChatGPT. Then we generate diverse prompts with random combinations of the attributes, which are in turn used to replace the simple class-conditional prompt for querying the data from the ChatGPT. We empirically evaluate the generated datasets via the performance of the model trained with them in two scenarios: 1) train a model solely on the generated dataset, and 2) train a model on a dataset merged from the real training set and the generated set. On both scenarios, the dataset generated with diversely attributed prompts outperforms its counterpart with simple class conditional prompts by a large margin, demonstrating the superiority of the attributed prompts.

We then provide a comprehensive study on LLM as an attributed data generator and the applications of the generated datasets. We refer to the real training set, the dataset generated by simple class-conditional prompts, and that by attributed prompts as Gold, SimPrompt, and AttrPrompt. We aim to answer the following questions with empirical evidence:

[*]Equal contribution [1]Georgia Tech [2]University of Washington [3]Google Tech [4]UIUC. Correspondence to: Yue Yu <yueyu@gatech.edu>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. 

*Table 1.* Statistics of datasets.

| Dataset | Domain | Task | # Train | # Valid | # Test | # Class | Imbalance Ratio |
|---|---|---|---|---|---|---|---|
| Amazon (Blitzer et al., 2007) | Reviews | Multi-class | 15.0K | 0.2K | 1.2K | 23 | 155.6 |
| NYT (Meng et al., 2019) | News | Multi-class | 9.0K | 0.2K | 1.2K | 26 | 357.0 |
| Reddit (Geigle et al., 2021) | Web | Multi-class | 26.6K | 0.2K | 2.3K | 45 | 447.4 |
| StackExchange (Geigle et al., 2021) | Web | Multi-class | 27.0K | 0.3K | 2.5K | 50 | 1283.7 |

**Is there systematic bias in real and generated data?** We leverage the AttrPrompt dataset and the attributes associated with each data to train an attribute classifier, serving as a probe to uncover the systematic bias behind both the Gold and the SimPrompt dataset. For the "*location*" attribute of the NYT news dataset, the predicted values of both Gold and SimPrompt are dominated by "*North America*", while "*Africa*" is extremely rare. Such a regional bias in both real and generated dataset could be a roadblock for building trustworthy machine learning models.

**How important the attribute diversity is?** We examine the impact of the attribute diversity on the model performance. We first identify the best attribute value for each attribute dimension by one-fixed-others-random strategy, and then show that composing all the individually best attribute values, which has zero attribute diversity, leads to significantly worse performance than random attribute configurations. This emphasizes the importance of the attribute diversity and pose a new challenge of searching for optimal attributes for training data generation.

In addition, we study the budget and sample efficiency of the generated datasets, concluding AttrPrompt enjoys better efficiency than SimPrompt. We also investigate how well the AttrPrompt performs with long-tail classes and the effect of the temperature parameter on the generated dataset.

## 2 Related Work

With the remarkable success of large language models (LLMs), researchers have recently attempted to leverage them as the training data generators. Such applications include generating tabular data (Borisov et al., 2023), medical dialogue (Chintagunta et al., 2021), sentence pairs (Schick & Schütze, 2021), instruction data (Wu et al., 2023; Peng et al., 2023), *etc.*. Among these applications, we anchor on training data generation for text topic classification in a zero-shot setting where no labeled data is available. In this direction, existing approaches typically use simple class-conditional prompts while focusing on mitigating low-quality issues after generation. In particular, ZeroGen (Ye et al., 2022a) took an initial step to explore using LLM as a training data generator for topic classification with simple class-conditional prompts; to deal with the low-quality issue of the generated data, SuperGen (Meng et al., 2022) adopts label smoothing and temporal ensembling, SunGen (Gao et al., 2023) reweights the generated data during training with learned data quality weight, and ProGen (Ye et al., 2022b) leverages the model feedback to select highly influential generated data which then serve as labeled examples for generation. In this work, we instead explore attributed prompts to reduce the issue of low informativeness and redundancy, which can be readily incorporated into the existing systems mentioned above. Notably, (Chen et al., 2023) also explores prompts to advance the data generation process, yet it adopts soft prompts and requires a white-box LLM and seed examples to tune them. In contrast, our method is applicable to black-box LLMs and even LLM APIs (*e.g.*, ChatGPT) and does not rely on labeled examples.

## 3 Large Language Model as Attributed Training Data Generator

### 3.1 Datasets

While existing works either only involve binary classification datasets (Ye et al., 2022a; Meng et al., 2022; Ye et al., 2022b) or use datasets with at most 14 classes (Gao et al., 2023), it is unclear how LLM performs as data generator for text topic classification with high cardinality (many topic classes). Thus, in this work, we consider the following datasets from various domains with number of topics ranging from 23 to 50:

- **NYT** (Meng et al., 2019): The NYT dataset comprises news articles that were authored and published by The New York Times. These articles are categorized into 5 broad genres (such as arts and sports) and 26 fine-grained categories.
- **Amazon** (Blitzer et al., 2007): The Amazon dataset contains customer reviews on products from Amazon's online store. It covers products from 23 different categories.
- **Reddit** (Geigle et al., 2021): The Reddit dataset consists of a vast collection of user-generated content from the popular social media platform Reddit. It encompasses a wide range of topics, discussions, and interactions among users across numerous communities.
- **StackOverflow** (Geigle et al., 2021): The StackExchange dataset is a rich collection of structured data encompassing various online communities and knowledge-

*Table 2.* Attribute dimensions and values. Attributes with asterisk are class-dependent attributes.

| Dataset | # configurations / class | Attribute dimension | Attribute value |
|---|---|---|---|
| NYT | 600 | Subtopic* | Appendix B.1.1 |
| | | Location | Asia, North America, South America, Africa, Oceania, Europe |
| | | Writing Style | Investigative journalism, Op-Eds, Feature writing, News analysis, Profiles and interviews |
| | | Length | short (30-80 words); long (100-150 words) |
| Amazon | 1000 | Product Brands* | Appendix B.2.1 |
| | | Product Names* | Appendix B.2.2 |
| | | Usage Experience | Worst, Bad, Average, Good, Excellent |
| | | Writing Style | Detailed Review; Comparative Review; Pros and Cons Review; Recommendation Review |
| | | Length | short (30-80 words); long (100-150 words) |
| Reddit | 500 | Resources* | Appendix B.3.1 |
| | | Experience* | Appendix B.3.2 |
| | | Writing Style | Informative/Educational; Entertaining/Funny; Discussion; Storytelling; Help/Advice |
| | | Length | short (30-80 words); long (100-150 words) |
| StackExchange | 400 | Scenario* | Appendix B.4.1 |
| | | Technical Depth | Beginner; Intermediate; Advanced; Expert |
| | | Writing Style | Specific; Comparative; Problem-Solution; Troubleshooting; Tutorial |
| | | Length | short (30-80 words); long (100-150 words) |

*Table 3.* Prompt template.

| Method | Prompt |
|---|---|
| SimPrompt | Suppose you are a news writer. Please generate a `{topic}` news in NYT. |
| AttrPrompt | Suppose you are a news writer. Please generate a `{topic}` news in NYT following the requirements below:<br>1. Should focus on `{subtopic}`;<br>2. Should be in length between `{length:min-words}` and `{length:max-words}` words;<br>3. The writing style of the news should be `{style}`;<br>4. The location of the news should be in `{location}`;<br>5. The news must be irrelevant to `{similar-classes}`. |

sharing platforms. It contains a vast array of questions, answers, comments, tags, and user interactions about specific technical problems.

For Reddit and StackOverflow, we select the classes with more than 65 examples from the original corpus as the target set of topics. For each dataset, we use 50 examples per class for the test set and no more than 10 examples for the validation set (10 for NYT/Amazon and 5 for Reddit/StackOverflow). The remaining data is used to compose the gold training set. It is worth noting that, some of the class names on Reddit may contain toxic information. To eliminate their effects, we filter our label names with Detoxify (Hanu & Unitary team, 2020), a tool that leverages the multilingual XLM-RoBERTa (Conneau et al., 2019) for toxic comment identification. We follow (Gadre et al., 2023) to use a threshold of 0.1 to filter out potentially toxic topic classes. We summarize the statistics of used dataset in Table 1, from which we can see that the involved datasets not only have high cardinality but also come with high imbalance ratio, *i.e.*, the ratio of sample size of the majority class to that of the minority class, which reflects the long-tail class issue in real applications.

### 3.2 Attributes

Our first step is identifying various kinds of data attributes (or metadata) that one could manipulate to generate attributed data samples. In particular, data attributes consist of attribute dimension and corresponding attribute values, where the latter are possible instantiations of the former. For example, attribute value "*shorter than 200 words*" could be an instantiation of attribute dimension "*length*". In this work, we adopt a human-AI collaboration paradigm (Liu et al., 2022) to create both attribute dimensions and attribute values. Specifically, we start by querying ChatGPT to generate crucial attribute dimensions. To accomplish this, we pose questions such as "*Which attribute dimensions do you consider vital in determining the topic of a news article?*" for the NYT dataset and obtain answers like "*subtopics, length, style, location, reader group, style, time*". Then, we manually select the high-quality attribute dimensions that best suit the dataset (*e.g.*, "*location*", "*subtopics*", *etc.*). Similarly, we prompt ChatGPT for potential attribute values within each attribute dimension and carefully select high-quality candidates.

*Table 4.* Data examples of different datasets: the "*federal budget*" class of the NYT dataset.

| Method | Example |
|---|---|
| **Gold** | 1. The emergency manager who was sent to reverse the fortunes of this financially troubled city asked some of its creditors on Friday to accept pennies on the dollar as he laid out his plan for tackling Detroit's staggering debt, kick starting negotiations that could determine whether the city is headed to bankruptcy court...<br>2. Saying that its debt could reach 45 billion by 2017 if Congress does not act, the Postal Service on Wednesday called on lawmakers to give it the flexibility to change its business model to keep itself solvent.During a hearing before the Senate Homeland Security and<br>3. Governmental Affairs Committee, Patrick R. Donahoe, the postmaster general, asked Congress to give the Postal Service permission to run its own health plan for employees and retirees, modify a Congressional mandate that requires the agency to pay...<br>4. Voters want things. And politicians officeholders, especially are responsive to voters .Cranky people stuck in traffic are not happy people, and they pass their unhappiness along to those who represent them. Those same voters, in many cases, have already expressed unhappiness to those responsive and able officeholders about some other things, too things that, in politics, might get somebody's mouth washed out with soap taxes, debt and tolls...<br>5. SACRAMENTO California has been Exhibit A for the fiscal upheaval that has rocked states throughout the recession . Year after year, California officials reported bigger and bigger deficits and sought to respond with spending cuts that left the state reeling... |
| SimPrompt | 1. Washington D.C. The United States government has recently passed its federal budget for the next fiscal year, setting aside a whopping 4.8 trillion for spending. This is the largest budget in the country's history and reflects the government's commitment to improving the country's economy and infrastructural development.<br>2. WASHINGTON D.C. The Trump administration released its proposed budget for the federal government on Monday, calling for cuts to programs that provide assistance to low income Americans while increasing military spending. The proposed budget includes a 10 percent increase in defense spending, bringing the total military budget to 574 billion...<br>3. Washington D.C. The U.S. federal government has presented its budget for the fiscal year 2022, which includes a total of 6 trillion in spending and initiatives aimed at making the country more competitive and sustainable. The budget plan includes significant investments in infrastructure, healthcare, education, and renewable energy.<br>4. The Senate approves a 1.9 trillion federal budget proposal for 2021 In a partisan vote, the Senate has approved a 1.9 trillion federal budget proposal that includes funding for COVID 19 relief, infrastructure, and social programs. Democrats supported the budget, arguing that it will help hard hit Americans recover from the pandemic, while Republicans criticized it for being too expensive and lacking transparency... |
| AttrPrompt | 1. In Germany, public opinion on federal budget issues has been influenced by The Federal Agency for Civic Education. Using in depth storytelling, this agency communicates budget policies to citizens, emphasizing their impact on society. Its efforts have led to greater engagement and awareness of budget issues among the public in recent years.<br>2. Op Eds The Government Spending Priorities in Mexico As the Mexican federal budget draws near, all eyes are on the government's spending priorities. One area of concern is the role of the Banco de Mexico in managing the country's finances. Many experts argue that the central bank should play a more active role in balancing the budget and stimulating economic growth...<br>3. TOKYO The recent federal budget proposal has triggered a wave of agency budget cuts and restructuring in Japan, with the Japan Foundation for Regional Development being one of the latest casualties. The foundation, which aims to promote regional development and revitalization, is set to have its budget slashed by 20 next year.<br>4. In Colombia's federal budget, tax reform proposals gained traction, with the Administrative Tribunal of Cundinamarca being a vocal supporter. In an interview, the tribunal's president praised the potential to increase revenues without hurting the economy.<br>5. The South African government's fiscal policy decisions are drawing attention to the Independent Regulatory Board for Auditors (IRBA) after the Finance Minister Tito Mboweni announced budgetary cuts. The cuts, imposed on all government departments, have led to a decline in IRBA's funding and raised concerns about the quality of audits conducted by auditors... |

**Attribute dimensions and values.** There are two types of attribute dimension, *i.e.*, class-independent attribute and class-dependent attribute, where the former is invariant to the classes of a dataset (*e.g.*, "*length*") while the latter is class-dependent, *e.g.*, different classes would have different attribute values for an attribute dimension of "*subtopic*". For class-independent attributes, we produce attribute values shared by all the classes, while for class-dependent attributes, we generate attribute values separately with the assistance of ChatGPT. We list attribute dimensions and values for all datasets in Table 2. These data attributes offer a human-manipulatable interface for attributed data generation. In this study, we examine the potential of leveraging attributes to improve the data generation process when using LLM as data generator, while leaving the search for the optimal data attributes for a specific task to future work.

**Class-dependent attribute value filtering.** For class-dependent attributes, it is important to let its attribute values be related to the associated class only, otherwise the generated sample could be ambiguous and possibly related to multiple classes. For example, for the "*economy*" class of the NYT dataset, one candidate attribute value of the "*subtopic*" produced by ChatGPT is "*effect of trade tariffs on manufacturing companies*", which is also related to another class of "*international business*" in the NYT data and thus may result in the generated data being ambiguous. To filter out ambiguous class-dependent attribute values of a given class, we first query the ChatGPT for its top-5 similar classes and then for each class-dependent attribute value, we ask the ChatGPT whether the value is related to the top-5 similar classes. If the answer is yes, we then remove this attribute value for the given class. We refer to this filtering as class-dependent attribute value filtering, shorten by CAF.

### 3.3 Prompts

Given data attributes, one could prompt LLM to generate data samples with diverse *attribute configurations*. For example, for the NYT dataset, an attribute configuration for the "*federal budget*" class could be {"*subtopic*"="*defense spending*","*length*"="*short:min-words=30,max-words=80*", "*style*"="*investigative journalism*", "*location*"="*North America*"}. In Table 2, we list the number of configurations per class, one can further enlarge the number of configurations by adding more attribute dimensions and values. To generate attributed data samples, we prompt ChatGPT with random configurations. In particular, each time we generate a random configuration, complete a *prompt template* with the generated configura-

*Table 5.* Performance of the models trained with created datasets and the cost of constructing the datasets. We additionally include the performance and cost of using LLM as a zero-shot predictor.

| Method | NYT | | | Amazon | | | Reddit | | | StackExchange | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | F1 | Price/1k | Acc. | F1 | Price/1k | Acc. | F1 | Price/1k | Acc | F1 | Price/1k |
| LLM Zero-Shot | 74.16 | 69.84 | 5.44 | 59.55 | 54.56 | 2.11 | 67.00 | 56.66 | 2.89 | 44.70 | 43.80 | 3.12 |
| Gold | 83.80 | 81.02 | — | 82.23 | 81.12 | — | 84.22 | 83.38 | — | 67.56 | 63.28 | — |
| SimPrompt | 75.47 | 76.22 | 0.76 | 57.34 | 56.96 | 0.77 | 53.48 | 53.81 | 0.65 | 42.88 | 41.30 | 0.69 |
| AttrPrompt w/o CAF | 80.40 | 80.92 | 0.91 | 61.67 | 61.57 | 0.82 | 61.22 | 60.18 | 0.72 | 45.90 | 44.84 | 0.81 |
| AttrPrompt | 81.30 | 82.26 | 1.05 | 66.28 | 65.87 | 0.87 | 63.33 | 63.10 | 0.84 | 48.99 | 47.42 | 0.90 |

*Table 6.* Performance of the models trained with the original training set augmented with the generated dataset. We also present the performance gain/drop compared to using the original training set only in color.

| Method | NYT | | Amazon | | Reddit | | StackExchange | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | F1 | Acc. | F1 | Acc. | F1 | Acc | F1 |
| SimPrompt | 85.56 +1.76 | 86.34 +5.32 | 81.85 -0.38 | 80.23 -0.89 | 85.11 +0.89 | 84.88 +1.50 | 74.53 +6.97 | 74.23 +10.95 |
| AttrPrompt w/o CAF | 85.71 +1.91 | 87.18 +6.16 | 82.24 +0.01 | 80.76 -0.36 | 85.86 +1.64 | 85.65 +2.27 | 75.16 +7.60 | 74.64 +11.36 |
| AttrPrompt | 87.47 +3.67 | 88.06 +7.04 | 83.95 +1.72 | 83.93 +2.81 | 86.08 +1.86 | 85.98 +2.60 | 76.86 +9.30 | 76.53 +13.25 |

tion, and query the ChatGPT with the completed prompt to collect generated data samples. We present the used prompt template (denoted as AttrPrompt) in Table 3 as well as a simple class-conditional prompt (denoted as SimPrompt) without using data attributes as a comparison.

# 4 Experiments

In this section, we compare our method (AttrPrompt) against simple class-conditional prompt (SimPrompt) and the original training set of each dataset (Gold). For a fair comparison, we set the number of generated data the same as Gold for both AttrPrompt and SimPrompt. In principle, the generated dataset can be used to train any classifier; if not otherwise specified, we choose to fine-tune `BERT-base-uncased` (Devlin et al., 2019) as the backbone.

## 4.1 A Glimpse of the Generated Data

*Table 7.* Comparison of the vocabulary size of different datasets.

| Method | NYT | | Amazon | | Reddit | | StackExchange | |
|---|---|---|---|---|---|---|---|---|
| | All | Class Avg. | All | Class Avg. | All | Class Avg. | All | Class Avg. |
| Gold | 70.8k | 11.3k | 44.7k | 6.64k | 50.8k | 4.62k | 52.3k | 3.60k |
| SimPrompt | 20.6k | 3.13k | 11.6k | 2.50k | 19.9k | 3.06k | 13.3k | 2.20k |
| AttrPrompt | 21.4k | 3.50k | 14.0k | 2.76k | 25.4k | 3.64k | 17.8k | 2.93k |

We first show several examples of data generated by AttrPrompt and SimPrompt and real data in Gold for the "*federal budget*" class of the NYT dataset (Table 4). We can see that the data generated by the ChatGPT have high-quality. Notably, compared with SimPrompt, AttrPrompt renders more diverse samples since SimPrompt tends to generate news about the U.S. while AttrPrompt could produce news all over the world.

Then, we show the vocabulary size of the generated dataset and the Gold dataset, which is a natural way to check the lexical diversity of datasets (Table 7). From the table, we can see that AttrPrompt has higher lexical diversity than SimPrompt in terms of both vocabulary size of the whole dataset (All in the table) and the averaged vocabulary size across classes (Class Avg. in the table). Yet, both have much less vocabulary size than the Gold, indicating there is still room for improvement of the ChatGPT's lexical diversity.

## 4.2 Training with generated data

We quantitatively evaluate the quality of generated datasets via the test performance of models trained with them. In addition, we use the ChatGPT as a zero-shot predictor for comparison. The results are in Table 5. Besides the test performance, we include the cost of querying the ChatGPT per 1000 data in the table.

From the results, we can draw the following conclusions. First, the AttrPrompt consistently renders better performance compared to the SimPrompt with a margin of 6–10 points. Second, the class-dependent attribute value filter (CAF) is beneficial since the AttrPrompt outperforms its variant without CAF. Third, out of the four datasets, the AttrPrompt outperforms the LLM zero-shot method on three datasets in terms of accuracy, while for the F1 score, the AttrPrompt surpasses the LLM zero-shot on all the datasets; combined with the observation that the LLM zero-shot inference incurs much higher costs compared to data generation and the fact that the generated data is re-usable for training any model, we argue that for topic text classification generating training data could be a better practice of leveraging LLM than direct zero-shot inference. Lastly, in most cases, the generated data underperform the original training set,


(a) SimPrompt: All. (b) SimPrompt: "*tennis*". (c) SimPrompt: "*soccer*". (d) SimPrompt "*international business*".

(e) Gold: All. (f) Gold: "*tennis*". (g) Gold: "*soccer*". (h) Gold: "*international business*".

*Figure 1.* Pie charts of the distributions of "*location*" predicted by an attribute classifier for the NYT SimPrompt and Gold dataset. (a) and (e) are "*location*" distribution over the whole dataset, while others are for specific classes.

indicating that there is still room for future improvement.

### 4.3 Augmenting existing dataset with generated data

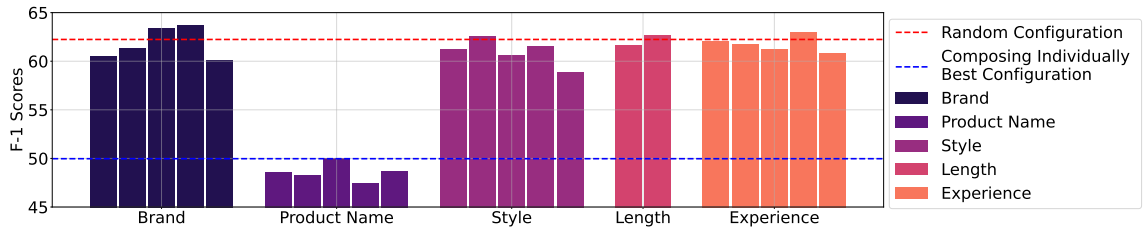
Here, we merge the generated dataset and the original training set into a single training set, and then test the model performance when it is trained with the merged dataset to see whether the generated dataset can further improve model performance with the original training set available. We present the results in Table 6. From the table, we can see that the generated dataset is an effective complement to the original training set, since most of the generated datasets introduce performance gain when combined with the original training set, especially our AttrPrompt which leads to improvement for all the cases. This improvement with simple dataset merge may motivate future studies of more advanced ways of using the generated data as augmentations to boost existing datasets.

## 5 Analysis

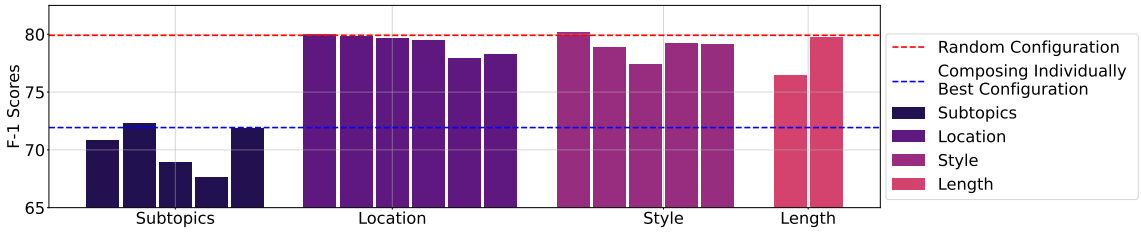
### 5.1 Is there systematic bias in real and generated data?

We study the systematic bias in both real dataset (Gold) and generated dataset of SimPrompt using dataset generated by AttrPrompt as a probe. In particular, we leverage the attributes associated with each data of AttrPrompt to train an *attribute classifier*, which is in turn used to make attribute predictions on Gold and SimPrompt dataset. Note that the attribute values associated with each data of AttrPrompt is not necessary the ground truth, yet since ChatGPT has shown remarkable performance in following instructions (), the generated data could decently reflect the desired attributes and therefore the attribute classifier trained with them could partially reveal the underlying attribute distribution of tested dataset, *i.e.*, Gold and SimPrompt.

We pick the "*location*" attribute of the NYT data and visualize the distributions of the predicted "*location*" of Gold and SimPrompt in pie charts (Figure 1). One can see that both the Gold and SimPrompt dataset are largely biased towards "*North America*" in terms of the whole dataset (subfigures (a)&(e) in Figure 1). As to the "*location*" distribution for specific classes, we can see that Gold and SimPrompt are biased towards continents other than "*North America*". In contrast, with attributed prompts, the generated dataset of AttrPrompt comes with relatively balanced attribution distribution. While existing work of using LLM as data generator usually overlook the bias embedded in the generated data, we hope that this preliminary analysis could raise the attention of the community to the systematic bias behind the generated data of powerful LLM such as ChatGPT.




(a) Amazon.




(b) NYT.

*Figure 2.* Bar chart comparison between single configuration and random composed configuration.

### 5.2 How important the attribute diversity is?

We study the effect of the attribute diversity of AttrPrompt on the model performance. In particular, for each attribute dimension, we fix its value to be one of the candidate values each time and keep the value of other attributes random. Then, we generate 50 data per class using such a one-fixed-others-random configuration to compose a dataset and evaluate the performance of the trained model. Note that for class-dependent attributes, we sample one value for each class and repeat it 5 times, since it is computationally prohibitive to enumerate all combinations of class-dependent attribute values. In Figure 2, each bar stands for a specific one-fixed-others-random configuration; compared to random configurations, most of one-fixed-others-random configurations result in a performance drop. To further reduce the attribute diversity, we pick the attribute value with the best performance for each attribute dimension (the highest bar within each attribute dimension) and compose them to a single configuration (the dashed blue line). We can see that the dashed blue line is significantly worse than the random configuration, even though it is composed of individually best attribute values. This illustrates the importance of prompts with diverse attributes.

### 5.3 The budget and sample efficiency

Here, we aim to study two types of efficiency of the generated dataset, *i.e.*, budget efficiency and sample efficiency, on the model performance. First, in Figure 4 we compare the budget efficiency of AttrPrompt against that of SimPrompt. Surprisingly, AttrPrompt only requires 5% of budget to be on par with or outperform SimPrompt with 100% of budget across all the datasets. This observation demonstrates the importance of diverse prompts in training data generation.

Secondly, we compare the sample efficiency of Gold, SimPrompt, and AttrPrompt in Figure 5. On one hand, compared with the Gold, both SimPrompt and AttrPrompt have better sample efficiency in low-data regime, since they outperform the Gold when the dataset size is relatively small; on the other hand, Gold data exhibits better sample efficiency at high-data regime. Overall, AttrPrompt renders better sample efficiency than SimPrompt, which suggests that increasing the diversity of the prompts could be an effective way to improve the unsatisfactory data scaling trend of using LLM as data generator (Ye et al., 2022b).

### 5.4 Can the generated data remedy the long-tail classes issue?

As we have seen in Table 1, the original training sets of the involved datasets have severe long-tail classes issue since the imbalance ratio is high, yet the generated dataset are class-balanced, we are then curious how the class balance in the generated dataset benefits the model performance on long-tail classes. We take the NYT dataset as an example and plot the per-class F1 score of Gold, SimPrompt, and AttrPrompt in Figure 3, where the x-axis is classes sorted by their number of data in the Gold dataset in descending order. From the figure, we can see that out of 26 classes, AttrPrompt renders the best per-class F1 score on 10 classes, which is 13 for Gold and 3 for SimPrompt. Notably, for classes with few examples in the Gold set (the rightmost 4 classes in the figure), AttrPrompt is better than the Gold and SimPrompt, especially for the class "*abortion*" with the fewest examples. This suggests a data-centric way to handle the long-tail class issue in topic classification: one may use

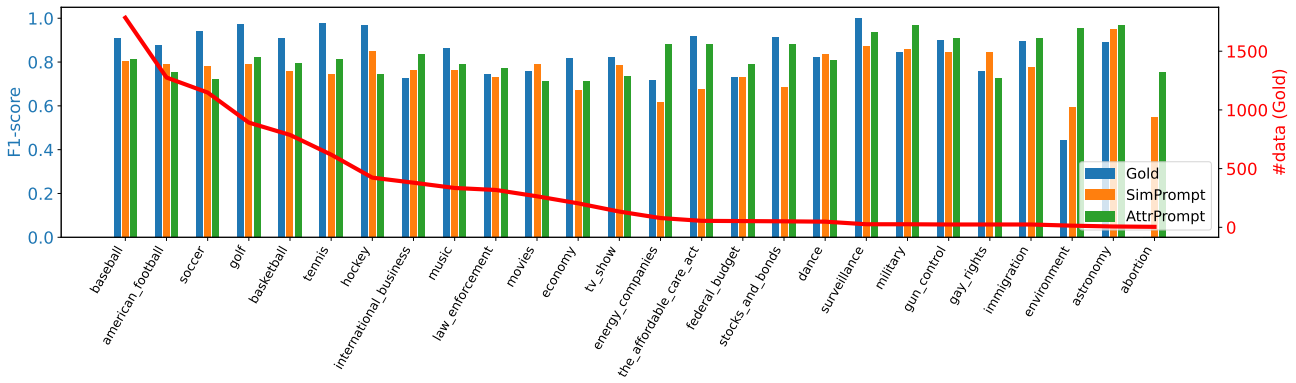
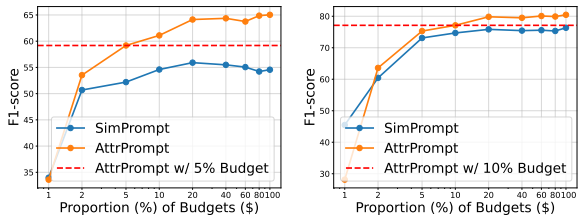
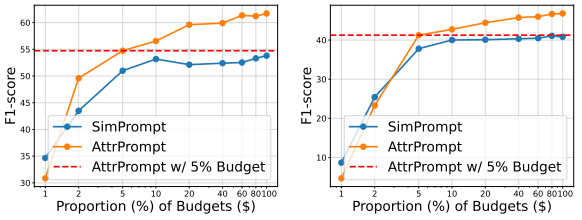



*Figure 3.* Per-class F1-score of the NYT dataset.

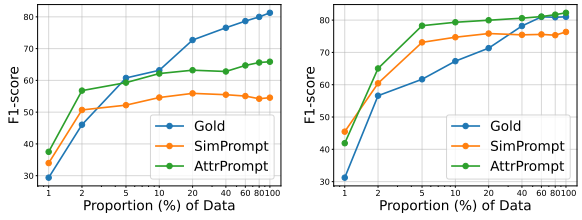



(a) Amazon (b) NYT

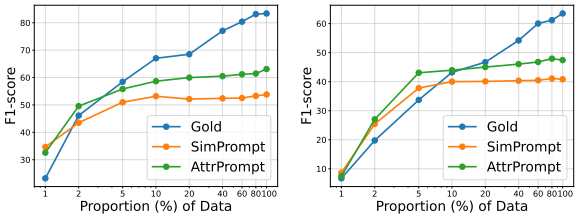



(c) Reddit (d) StackExchange

*Figure 4.* The comparisons on budget efficiency.




(a) Amazon (b) NYT




(c) Reddit (d) StackExchange

*Figure 5.* The comparisons on data efficiency.

LLM to generate class-balanced training set or augment existing training set with the LLM-generated data such that the augmented dataset is class-balanced, the in-depth study of which is left as future work.

### 5.5 Different temperature parameters for ChatGPT

Temperature ($t$) is one crucial hyperparameter of LLMs that controls the diversity of the generated text, while the studied attributed prompts are also for diversifying the generated data. We are then curious about the effectiveness of the temperature and how it compares to the AttrPrompt. We study different values of the temperature using the NYT dataset and present the results in Table 8. From the results, we can see that compared with the temperature, AttrPrompt brings more significant performance gain, demonstrating its superiority over temperature tuning.

*Table 8.* Study of the temperature.

| **Method** | $t = 1.0$ | | $t = 1.5$ | | $t = 2.0$ | |
|---|---|---|---|---|---|---|
| | Acc. | F1 | Acc. | F1 | Acc. | F1 |
| SimPrompt | 76.00 | 76.34 | 76.78 | 77.31 | 76.55 | 77.42 |
| AttrPrompt | 81.30 | 82.26 | 81.47 | 82.23 | 79.47 | 79.86 |

## 6 Conclusion

We delve into the realm of training data generation using complex, attributed prompts, which possess the potential to produce a wide range of diverse and attributed generated data. Specifically, we focus on datasets characterized by high cardinality and diverse domains, and our results demonstrate the superior performance of attributed prompts compared to simple class-conditional prompts. Furthermore, we present a comprehensive empirical study on training data generation that covers essential aspects such as bias, diversity, and efficiency.

## References


Blitzer, J., Dredze, M., and Pereira, F. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 440–447, 2007.

Borisov, V., Sessler, K., Leemann, T., Pawelczyk, M., and Kasneci, G. Language models are realistic tabular data generators. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cEygmQNOeI.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chen, D., Lee, C., Lu, Y.-Y., Rosati, D., and Yu, Z. Mixture of soft prompts for controllable data generation. *ArXiv*, abs/2303.01580, 2023.

Chintagunta, B., Katariya, N., Amatriain, X., and Kannan, A. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, pp. 354–372. PMLR, 2021.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Ferrara, E. Should chatgpt be biased? challenges and risks of bias in large language models. *ArXiv*, abs/2304.03738, 2023.

Gadre, S. Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023.

Gao, J., Pi, R., Yong, L., Xu, H., Ye, J., Wu, Z., Zhang, W., Liang, X., Li, Z., and Kong, L. Self-guided noise-free data generation for efficient zero-shot learning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=h5OpjGd_lo6.

Geigle, G., Reimers, N., Rücklé, A., and Gurevych, I. Tweac: transformer with extendable qa agent classifiers. *arXiv preprint arXiv:2104.07081*, 2021.

Hanu, L. and Unitary team. Detoxify, 2020. https://github.com/unitaryai/detoxify.

He, P., Gao, J., and Chen, W. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. Tinybert: Distilling bert for natural language understanding. In *Findings of EMNLP*, pp. 4163–4174, 2020.

Kirk, H. R., Jun, Y., Iqbal, H., Benussi, E., Volpin, F., Dreyer, F. A., Shtedritski, A., and Asano, Y. M. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models. In *Neural Information Processing Systems*, 2021.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Liu, A., Swayamdipta, S., Smith, N. A., and Choi, Y. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of EMNLP*, pp. 6826–6847, 2022.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Meng, Y., Shen, J., Zhang, C., and Han, J. Weakly-supervised hierarchical text classification. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 6826–6833, 2019.

Meng, Y., Huang, J., Zhang, Y., and Han, J. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=4G1Sfp_1sz7.

OpenAI. Introducing chatgpt, 2023a. URL https://openai.com/blog/chatgpt.

OpenAI. Gpt-4 technical report. *arXiv*, 2023b.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Peng, B., Li, C., He, P., Galley, M., and Gao, J. Instruction tuning with gpt-4. *ArXiv*, abs/2304.03277, 2023.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Schick, T. and Schütze, H. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6943–6951, 2021.

Tevet, G. and Berant, J. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 326–346, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.25. URL https://aclanthology.org/2021.eacl-main.25.

Wu, M., Waheed, A., Zhang, C., Abdul-Mageed, M., and Aji, A. F. Lamini-lm: A diverse herd of distilled models from large-scale instructions. 2023.

Ye, J., Gao, J., Li, Q., Xu, H., Feng, J., Wu, Z., Yu, T., and Kong, L. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11653–11669, 2022a. URL https://aclanthology.org/2022.emnlp-main.801.

Ye, J., Gao, J., Wu, Z., Feng, J., Yu, T., and Kong, L. ProGen: Progressive zero-shot dataset generation via in-context feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3671–3683, 2022b. URL https://aclanthology.org/2022.findings-emnlp.269.

Yu, Y., Zhuang, Y., Zhang, R., Meng, Y., Shen, J., and Zhang, C. Regen: Zero-shot text classification via training data generation with progressive dense retrieval. In *Findings of ACL*, 2023.

Zhuo, T. Y., Huang, Y., Chen, C., and Xing, Z. Exploring ai ethics of chatgpt: A diagnostic analysis. *ArXiv*, abs/2301.12867, 2023.

# A Implementation Details

## A.1 Hardware Information

All experiments are conducted on *CPU*: Intel(R) Core(TM) i7-5930K CPU @ 3.50GHz and *GPU*: NVIDIA GeForce RTX A5000 GPUs using python 3.8, Huggingface 4.6.0 and Pytorch 1.10.

## A.2 Parameter Configuration

We keep the parameter $\text{top\_p} = 1.0$ and temperature $t = 1.0$ for calling ChatGPT APIs (OpenAI, 2023a) for the training data generation part. For finetuning the classifier, we optimize the model using AdamW (Loshchilov & Hutter, 2019) with a linear warmup of the first 5% steps and linear learning rate decay. The maximum number of tokens per sequence is 128. Table 9 lists the hyperparameters used for AttrPrompt and SimPrompt. For the generated synthetic dataset, we stick to the strict zero-shot learning setting (Meng et al., 2022), train all the models for 6 epochs and use the model from the last epoch *without using the validation set* for evaluation. For the original clean dataset, we train all models for 6 epochs and use the model with the best performance on the validation set for evaluation.

| Backbone | Learning Rate lr | Batch Size | Training Epochs $E$ | Weight Decay | Warmup Ratio |
|---|---|---|---|---|---|
| `BERT-base-uncased` (Devlin et al., 2019) | 5e-5 | 32 | 6 | 1e-4 | 5% |
| `TinyBERT` (Jiao et al., 2020) | 1e-4 | 32 | 6 | 1e-4 | 5% |
| `DistilBERT-base-uncased` (Sanh et al., 2019) | 5e-5 | 32 | 6 | 1e-4 | 5% |
| `DeBERTa-base-V3` (He et al., 2023) | 5e-5 | 32 | 6 | 1e-4 | 5% |
| `DeBERTa-large-V3` (He et al., 2023) | 2e-5 | 32 | 6 | 1e-4 | 5% |

*Table 9.* Hyperparameters for fine-tuning classifiers on different tasks.

# B Attribute Details

## B.1 NYT

### B.1.1 SUBTOPICS

We randomly select 5 categories in NYT dataset and display the corresponding subtopic attributes for each category:

- astronomy:
  - Discoveries of exoplanets
  - Black holes and their role in shaping galaxies
  - The search for extraterrestrial life
  - Gravitational waves and the study of the universe's origins
  - The use of telescopes to explore the universe
  - The mysteries of dark matter and dark energy
  - Solar flares and their impact on Earth
  - The history of the universe and its evolution over time
  - Exploring the possibility of space tourism
  - The exploration of our neighboring planets, such as Mars and Venus.
- baseball:
  - Recent controversy surrounding sign-stealing scandal in MLB
  - Breakdown of top prospects in minor league baseball
  - Analysis of new rule changes for upcoming baseball season
  - Coverage of recent World Series champions and their success
  - In-depth profile of influential baseball figures, such as managers or players
  - Updates on retired players and their post-baseball careers

  - Highlighting standout performances by individual players or teams in recent games
  - Coverage of international baseball leagues and their top players

- economy:
  - Job market and employment rates
  - Interest rates and monetary policy
  - Inflation and deflation
  - Economic growth and GDP
  - Consumer spending and retail sales
  - Income inequality and poverty
  - GDP growth and contraction
  - Labor market trends
  - Economic impacts of natural disasters and pandemics
  - Housing market and real estate

- federal_budget:
  - Deficit reduction strategies
  - Government spending priorities
  - Tax reform proposals
  - Agency budget cuts and restructuring
  - Budget negotiations and debates
  - National debt projections
  - Fiscal policy decisions
  - Congressional budget proposals
  - Infrastructure spending plans
  - Public opinion on federal budget issues

- movies:
  - Box office records and trends for Hollywood blockbusters
  - Pioneering techniques in film-making and special effects
  - Representation and diversity in casting and storytelling
  - Reviews and analysis of highly-anticipated new releases
  - The impact of streaming services on movie distribution and consumption
  - The intersection of politics and film, from socially-conscious storytelling to politically-charged controversies
  - Profiles of notable actors, directors, and producers shaping the industry
  - The changing landscape of film festivals and awards season
  - Spotlight on independent and international cinema
  - The legacy of classic films and their enduring cultural impact.

## B.2 Amazon

### B.2.1 Product Brands

We randomly select 5 categories in Amazon dataset and display the corresponding product brand attributes for each category:

- camera_photo.:
  - Canon
  - Nikon
  - Leica

  - Hasselblad
  - Fujifilm
  - Lumix (Panasonic)
  - GoPro
  - Polaroid
  - Yashica
  - Mamiya
- jewelry_and_watches.:
  - Rolex
  - Cartier
  - Tiffany & Co.
  - Bulgari
  - Omega
  - Patek Philippe
  - Swarovski
  - Gucci
  - Michael Kors
  - Pandora
- magazines.:
  - Rolling Stone
  - Vogue
  - National Geographic
  - The New Yorker
  - GQ
  - Cosmopolitan
  - People
  - Time
  - Sports Illustrated
  - Forbes
- health_and_personal_care.:
  - Johnson & Johnson
  - Dove
  - Colgate
  - Head  Shoulders
  - Merck
  - Pfizer
  - Merck
  - Abbott Laboratories
  - GlaxoSmithKline
- toys_games.:
  - Mattel
  - Fisher-Price
  - Hasbro
  - Lego

  - Nerf
  - Barbie
  - Hot Wheels
  - Playmobil
  - MGA Entertainment
  - Paw Patrol

#### B.2.2 PRODUCT NAMES

We randomly select 5 categories in Amazon dataset and display the corresponding product name attributes for each category:

- sports_outdoors.:
  - Trekking poles
  - Kayak
  - Stand-up paddleboard
  - Treadmill
  - Bike
  - Yoga mat
  - Weightlifting gloves
  - Athletic training cones
  - Ab wheel
  - Resistance bands
  - Jump rope
  - Rollerskates
  - Boxing gloves
  - Basketball
  - Football
  - Golf clubs
  - Tennis racquet
- camera_photo.:
  - SnapShooter
  - FocusReady
  - ClickCapture
  - MemoriesMatter
  - FlashFinder
  - PicturePerfect
  - ShotSnap
  - VisionVibe
  - PixelPal
  - FreezeFrame
  - AngleAttack
  - SnapStash
  - FrameFlair
  - SmileSnaps
  - ImageImpact
  - ZoomZest
  - ClickCraze

  - PixelPassion
  - ShootSmart
  - CaptionChamp.

- grocery_and_gourmet_food.:
  - Nutella
  - Doritos
  - Hellmann's Mayonnaise
  - Campbell's Soup
  - Ritz Crackers
  - Quaker Oats
  - Ben  Jerry's Ice Cream
  - Tostitos Salsa
  - Goldfish Crackers
  - Red Bull Energy Drink
  - McCormick Spices
  - Crystal Light Drink Mix
  - Funyuns Onion Rings
  - Skippy Peanut Butter
  - Heinz Ketchup
  - Tabasco Hot Sauce
  - Hershey's Chocolate Syrup
  - Nescafe Coffee
  - Kraft Macaroni  Cheese
  - Gatorade Sports Drink

- baby.:
  - Baby Swing
  - Diaper Genie
  - Milk Warmer
  - Baby Carrier
  - Car Seat
  - Baby Monitor
  - Baby Food Maker
  - Nursery Glider
  - Changing Table
  - Baby Bouncer
  - Playpen
  - Teething Rings
  - Baby Wipes Dispenser
  - Baby Bath Tub
  - Bibs
  - Baby Blankets
  - Pacifier Clip
  - Baby Sling
  - Baby Napper
  - Moses Basket

- outdoor_living.:
  - Sunbrella
  - Weber
  - Keter
  - Trex
  - Solaira
  - Tropitone
  - Bimini Solutions
  - La-Z-Boy Outdoor
  - Suncast
  - Beltwood
  - Quikrete
  - Cosco Outdoor Living
  - Anova Furnishings
  - Duramax
  - US Polymers
  - Ostrich Chairs
  - Carefree of Colorado
  - Tuff Coat
  - Fire Sense
  - Heritage Patios.

### B.3 Reddit

#### B.3.1 Resources

We randomly select 5 categories in Reddit dataset and display the corresponding resource attributes for each category:

- buddy_crossing.:
  - Meetup.com - a website that allows you to find and join groups of individuals with similar interests in your local area, including hiking, book clubs, and social events.
  - The Buddy System: Understanding Mental Illness and Addiction - a book that explores the biology of addiction and provides a guide for friends and family members of individuals struggling with these issues.
  - Lynda.com - a subscription-based online learning platform that provides courses on a variety of subjects including computer programming, business, web design, and more.
  - Codecademy.com - an interactive online platform that teaches coding skills for free or through a subscription.
  - Khan Academy - a nonprofit organization that provides free online courses in a wide range of subjects including math, science, and humanities to learners of all ages.
  - Duolingo - a language-learning app that is available for free on the App Store and Google Play, offering courses in a variety of languages including Spanish, French, and German.
  - MindBody App - a mobile app that helps users find and book local fitness, wellness, and beauty activities such as yoga classes, massages, and haircuts.
  - Headspace - a meditation app that offers guided meditation courses to help users reduce stress and improve focus.
  - The Knot - a website that provides tools and resources for wedding planning, including a Wedding Website Builder, guest list tracker, and registry management.
  - Khan Academy - a nonprofit organization that provides free online courses in a wide range of subjects including math, science, and humanities to learners of all ages.
  - Others resource for buddy_crossing.
- the_division.:

  - Division Builds - A subreddit dedicated to sharing and discussing various builds used in The Division.
  - Division Zone - A website with extensive information on game mechanics, gear, and other important gameplay aspects.
  - The Division Discord - A community-run Discord server where players can connect and find groups to play with.
  - The Division Wiki - A comprehensive wiki with guides, tips, and information on everything related to The Division.
  - Skill-Up's YouTube channel - A popular YouTuber who provides detailed analysis and reviews of The Division's updates and patches.
  - MarcoStyle's YouTube channel - Another popular YouTuber who provides in-depth analysis and guides for The Division's gameplay and mechanics.
  - The Division LFG - A website where players can find groups to play with, organize events, and share their experiences.
  - The Division Zone Map - An interactive map that allows players to find important points of interest, loot, and other useful resources.
  - The Division 2 subreddit - A community-run subreddit for the sequel, The Division 2, where players can share their experiences and discuss the game.
  - Others resource for the_division.

- roblox.:
  - Roblox Wiki (https://roblox.fandom.com/wiki/Main_Page)
  - Roblox Developer Forum (https://devforum.roblox.com/)
  - Ultimate Guide to Making Your First Game on Roblox (https://medium.com/@Piranhari/ultimate-guide-to-making-your-first-game-on-roblox-part-1-f1fc63abfbe7)
  - Roblox Blog (https://blog.roblox.com/)
  - Roblox Studio Tutorials (https://www.youtube.com/playlist?list=PLuEQ5BB-Z1SgeZTAAq2w1K3kUfQ-yLEOj)
  - The Roblox Developer Hub (https://developer.roblox.com/en-us/)
  - Top 10 Roblox Games (https://www.techjunkie.com/top-10-best-roblox-games/)
  - Roblox Discord Server (https://discord.gg/roblox)
  - Roblox Support (https://en.help.roblox.com/hc/en-us)
  - Top Roblox Youtubers to Follow (https://www.gamertweak.com/top-roblox-youtubers-to-follow/)
  - Others resource for roblox.

- whats_that_book.:
  - Goodreads - A social platform for book lovers where users can search for books, create bookshelves, and write reviews.
  - LibraryThing - A community-driven cataloging website where users can create and share their personal book collections.
  - AbeBooks - An online marketplace for rare and out-of-print books, as well as other antique or collectible items.
  - Shelfari - An online book club where users can share book recommendations and read reviews from others.
  - Project Gutenberg - A digital library of freely available public domain books.
  - Paperback Swap - A book trading community where users can exchange books with others across the US.
  - Goodreads Librarians Group - A community of Goodreads users who help with book cataloging, including identifying books from incomplete information.
  - Book Riot - A website featuring book reviews and book-related news, with an emphasis on diverse and underrepresented voices.
  - The New York Times Book Review - A renowned weekly publication featuring book reviews, author interviews, and literary criticism.
  - Others resource for whats_that_book.

- pokemongo_friends.:

- Pokemon GO Hub: A comprehensive website dedicated to news, guides, and analysis on Pokemon GO.
- The Silph Road Subreddit: A community-run subreddit dedicated to research and analysis of Pokemon GO mechanics.
- Poke Assistant: A website that offers a range of tools to help you optimize your Pokemon GO experience, including IV calculators and gym battle simulations.
- The Trainer Club: A YouTube channel that provides daily updates, news, and tips for Pokemon GO trainers.
- Gotta Catch 'Em All: A Facebook group where you can connect with other Pokemon GO players and coordinate raid battles and other activities.
- Reddit's r/PokemonGOFriends Subreddit: A community of players looking for friends to exchange gifts and share invites for raids.
- The PokeMap: A website that allows you to find nearby Pokemon on a map in real-time.
- Poke Genie: An app that automatically calculates IVs and other stats for your Pokemon, saving you time and headaches.
- Pokemon GO Gamepress: A website that offers detailed breakdowns and analysis of Pokemon, movesets, and other game mechanics.
- The Go Ranger App: An app that helps you plan your raids and battles, with intuitive mapper tools and filters to help you find the Pokemon you're looking for.
- Others resource for pokemongo_friends.

#### B.3.2 EXPERIENCE

We randomly select 5 categories in Reddit dataset and display the corresponding experience attributes for each category:

- build_a_pc.:
  - DIY PC Builds: Sharing personal experiences and success stories of building custom PCs, discussing component choices, troubleshooting, and performance optimizations.
  - Budget-Friendly Builds: Discussing experiences with building PCs on a tight budget, sharing cost-saving tips, and recommendations for budget-friendly components.
  - Cable Management: Sharing personal experiences and tips for effective cable management in PC builds, discussing cable routing techniques and showcasing clean build aesthetics.
  - RGB Lighting: Discussing experiences with RGB lighting setups in PC builds, sharing recommendations for RGB components, software customization, and lighting effects.
  - Troubleshooting Builds: Sharing experiences and tips for troubleshooting common issues in PC builds, helping fellow builders diagnose and solve hardware or software problems.
  - Silent and Quiet PC Builds: Discussing experiences and recommendations for building silent or quiet PCs, focusing on noise reduction techniques and quiet component choices.
  - Workstation Builds: Sharing experiences and insights into building PCs for professional workloads, such as video editing, 3D rendering, programming, and graphic design.
  - Water-Cooling Adventures: Sharing experiences and insights into custom water-cooling loops, discussing the challenges, benefits, and performance improvements achieved.
  - Unique and Custom Builds: Showcasing and discussing unique and custom PC builds, including themed builds, custom cases, or exotic cooling solutions.
  - Build Planning and Component Selection: Discussing experiences with planning PC builds, researching and selecting components, considering compatibility, and balancing performance and budget.
  - Modding and Case Customization: Sharing experiences with PC case modding and customization, discussing techniques, materials, and showcasing personal projects.
  - Compact and Small Form Factor Builds: Discussing experiences with building compact or small form factor PCs, sharing recommendations for mini-ITX cases, cooling solutions, and component choices.
  - Home Server and NAS Builds: Sharing experiences and insights into building home servers and network-attached storage (NAS) systems, discussing storage options, software, and data management.

  - Multimonitor Setups: Discussing experiences with multimonitor setups, sharing tips for optimizing productivity and gaming experiences across multiple displays.
  - PC Gaming Peripherals: Sharing experiences and recommendations for gaming peripherals, such as keyboards, mice, monitors, and headsets, discussing features and personal preferences.

- summon_sign.:
  - Sunbro Covenant: Embracing the Sunbro covenant and assisting fellow players with jolly cooperation, earning sunlight medals and praising the sun together.
  - Fashion Souls: Sharing and showcasing unique and fashionable character builds, armor sets, and weapon combinations for aesthetic enjoyment.
  - Covenant Experiences: Sharing experiences and strategies related to various in-game covenants, such as the Darkwraiths, Blades of the Darkmoon, or Forest Hunters.
  - Community Creations: Showcasing community-created content, such as fan art, videos, or fan fiction, celebrating the creativity and talent within the Summon Sign community.
  - Lore-friendly Builds: Discussing and sharing character builds that are aligned with specific characters or factions within the game's lore, adding immersion and roleplaying elements.
  - Community Appreciation: Expressing gratitude and appreciation for the community, developers, and the overall enjoyment derived from the Dark Souls series and the cooperative multiplayer experiences.

- xbox.:
  - Xbox One exclusive games such as Halo 5, Forza Horizon 4, and Gears of War 4
  - Xbox One media and entertainment apps such as Netflix and Hulu
  - memorable gaming moments or achievements on the Xbox console.
  - Purchase Xbox One online.
  - Xbox Kinect motion sensor accessory
  - Xbox Play Anywhere program
  - Other Experience of Xbox

- pittsburgh.:
  - Visit the Andy Warhol Museum
  - Watch a Steelers football game at Heinz Field
  - Explore the Carnegie Museum of Natural History
  - Ride to the top of Mount Washington on the Duquesne Incline
  - Take a leisurely stroll through Phipps Conservatory and Botanical Gardens
  - Experience the history of the city at the Senator John Heinz History Center
  - Tour the University of Pittsburgh campus
  - Attend a performance at the Benedum Center for the Performing Arts
  - Take a walk along the Three Rivers Heritage Trail
  - Taste pierogies, kielbasa, and other traditional Pittsburgh foods
  - Admire the architecture of the Cathedral of Learning
  - Explore the Pittsburgh Zoo and PPG Aquarium
  - Gaze in awe at the exhibits in the Carnegie Science Center
  - Visit the National Aviary and get up close with tropical birds
  - Check out the local art scene on Penn Avenue in the Garfield neighborhood
  - Attend the Three Rivers Regatta, Pittsburgh's largest annual summer event
  - Take a bike ride on the Great Allegheny Passage trail
  - Ride the roller coasters at Kennywood Amusement Park
  - Discover the nightlife in the South Side neighborhood
  - Go shopping at the Strip District markets for locally-made goods and fresh produce.

- Others experience for pittsburgh.

- metal_gear_solid.:
  - Tactical Weapon Customization: Experimenting with various weapons, attachments, and equipment to tailor loadouts to different mission objectives and playstyles.
  - Character Development: Witnessing the growth and development of iconic characters such as Solid Snake, Big Boss, or Raiden throughout their respective story arcs.
  - Stealthy Takedowns: Executing silent and non-lethal takedowns, utilizing tranquilizer darts, chokeholds, or sneaking up on enemies from behind.
  - Tactical Planning: Strategizing mission approaches, analyzing enemy patrols, setting traps, and utilizing distractions to gain the upper hand.
  - Memorable Characters: Developing connections with unique and memorable characters, such as Otacon, Meryl, Revolver Ocelot, or The Boss.
  - Stealthy Gadgets: Utilizing gadgets and tools, such as the cardboard box, thermal goggles, or the iconic Solid Eye, to gain advantages during missions.
  - Emotional Storytelling: Experiencing emotionally impactful moments within the narrative, exploring themes of loss, betrayal, loyalty, and the human cost of warfare.
  - Espionage Tactics: Participating in undercover missions, gathering intelligence, infiltrating enemy bases, and sabotaging their operations.
  - Lore and Mythology: Delving into the intricate lore, conspiracies, and historical events within the Metal Gear Solid universe, including topics like The Patriots or the Philosopher's Legacy.
  - Groundbreaking Game Design: Appreciating the innovative gameplay mechanics, cinematic presentation, and attention to detail that have made the Metal Gear Solid series a beloved and influential franchise in the gaming industry.
  - Others experience for metal_gear_solid.

### B.4 StackExchange

#### B.4.1 SCENARIOS

We randomly select 5 categories in StackExchange dataset and display the corresponding scenario attributes for each category:

- multiplayer.:
  - Cheating/hacking in online games
  - Inappropriate player behavior
  - Unbalanced game mechanics
  - Difficulty connecting to multiplayer servers
  - Matchmaking errors
  - Unresponsive or laggy gameplay
  - Glitches in gameplay affecting online matches
  - Difficulty finding players to match with
  - Balancing player skills in matchmaking
  - Disconnects and dropped connections mid-game
  - Cross-platform compatibility issues
  - In-game communication problems
  - Difficulty managing and moderating game servers
  - Addressing griefing and trolling in multiplayer games
  - Managing player accounts and login systems
  - Implementing or improving anti-cheat measures

  - Community feedback and feature requests
  - Addressing game-breaking exploits
  - Ensuring fair and accurate reporting of player statistics
  - Addressing server crashes and downtime.

- terrain.:
  - Error in generative algorithms for creating terrain
  - Difficulty in implementing procedural terrain generation in a specific game engine
  - Inconsistencies in terrain generation across different devices
  - Issues with realism in terrain generation algorithms
  - Difficulty in implementing terrain physics and collision detection
  - Terrain rendering issues on low-spec hardware
  - Incompatibility between terrain generation and map or level editors
  - Optimization of terrain generation algorithms for speed and memory usage
  - Unwanted artifacts and glitches in terrain mesh generation
  - Compatibility issues between terrain generation algorithms and game engine systems
  - Difficulty in creating realistic terrain textures and materials
  - Inaccuracy of terrain elevation generation in certain geographic regions
  - Difficulty in implementing terrain deformation mechanics
  - Poor performance with large-scale terrain rendering and generation
  - Unwanted noise and roughness in generated terrain meshes
  - Compatibility issues between terrain generation and asset importation pipelines
  - Inaccuracy of terrain heightmap data due to low-quality input data sources
  - Difficulty in handling multi-layer terrain materials and textures
  - Poor performance with dynamic terrain generation and updates
  - Issues with biome and climate-based terrain generation.

- rendering.:
  - Difficulty creating realistic hair and fur in rendering software.
  - Debugging issues with transparent materials in a 3D rendering engine.
  - Crashes or slow performance when rendering large scenes in real time.
  - Trouble with anti-aliasing and other graphics optimization techniques.
  - Struggle with optimizing rendering quality on lower-end hardware.
  - Difficulty incorporating custom shaders into a game engine or rendering pipeline.
  - Figuring out how to use the latest rendering features in a legacy project.
  - Issues with rendering dynamic lighting in real time, without pre-baking.
  - Optimizing flicker or aliasing issues caused by fast-moving objects in a scene.
  - Solving glitches or crashes caused by malfunctioning GPU drivers.
  - Difficulty in rendering complex ocean or water simulations.
  - Troubleshooting issues with volumetric rendering in a 3D engine.
  - Finding the optimal rendering settings for a particular 3D model or scene.
  - Figuring out optimal texturing and lighting in a photorealistic rendering.
  - Creating procedural textures and materials in a game engine.
  - Debugging flicker issues caused by overlapping or intersecting geometry.
  - Difficulty in rendering realistic motion blur in 3D animation.
  - Solving imbalanced lighting in a photorealistic rendering of a room or scene.
  - Finding the optimal rendering settings for VR or AR applications.

  - Debugging issues with inaccurate or glitchy global illumination in a scene.

- procedural-generation.:
  - Improving the efficiency of procedural generation algorithms in Python.
  - Troubleshooting issues with memory usage in large-scale procedural generation projects.
  - Debugging issues with randomized content generation in procedural levels.
  - Implementing procedural generation techniques in C++ for game development.
  - Exploring the potential of machine learning in procedural content generation.
  - Optimizing the generation of complex 3D models using procedural algorithms.
  - Managing complexity and maintaining consistency in procedurally generated game worlds.
  - Addressing issues with procedural generation of text-based content, such as dialogue or item descriptions.
  - Developing tools to aid in the creation and testing of procedural generation algorithms.
  - Balancing the need for randomness with player expectations for fairness and balance.
  - Addressing issues with the procedural generation of music and sound effects.
  - Improving the visual quality of procedurally generated game assets.
  - Exploring ethical concerns around the use of AI in procedural content generation.
  - Developing procedural generation techniques for non-linear narratives or branching storylines.
  - Improving the procedural generation of inhabited environments, such as procedurally generated NPCs.
  - Addressing issues with the procedural generation of terrain features such as rivers and mountains.
  - Implementing procedural generation techniques for user-generated content.
  - Supporting multithreaded execution in procedural generation algorithms.
  - Ensuring procedural generation techniques are compatible with various game engines and frameworks.
  - Improving the scalability of procedural generation algorithms for use in multiplayer games.

- networking.:
  - Difficulty in troubleshooting network connection issues on a Linux system
  - Configuring a wireless access point for a large office space
  - Implementing load balancing across multiple servers in a network
  - Optimizing network throughput to reduce latency in a gaming environment
  - Implementing firewall rules to block unauthorized access to a network
  - Troubleshooting DNS resolution issues on a Windows server
  - Designing and implementing a secure VPN connection
  - Setting up a network file server for shared access among multiple clients
  - Configuring SNMP to monitor network traffic and utilization
  - Designing a network topology for a large enterprise with multiple locations
  - Troubleshooting issues with Ethernet switches in a data center environment
  - Implementing QoS to prioritize network traffic for critical applications
  - Configuring NAT and PAT to enable internet access for multiple devices
  - Setting up and configuring VLANs to segment a network
  - Troubleshooting issues with network printers in an office environment
  - Configuring routing protocols in a large network
  - Securing wireless access points to prevent unauthorized access
  - Troubleshooting issues with VPN connection stability and speed
  - Implementing network virtualization with virtual LANs and virtual switches
  - Designing and implementing an effective network security strategy to prevent data breaches.