

490 **A Limitation, Social Impact, and Future Work**

491 We acknowledge that AttrPrompt, while being effective on several tasks we studied, is still far from
492 being perfect. Here we give a honest discussion on the limitation, societal impacts, as well as potential
493 avenues for future works.

494 **A.1 Limitation**

495 One limitation of this research is that the investigation focuses primarily on text classification. While
496 this provides valuable insights into the performance of attributed prompts compared to simple class-
497 conditional prompts in these specific contexts, the generalizability of the findings to other data types
498 and tasks remains to be explored.

499 Besides, the proposed approach assumes the label name of the target task following prior works [31,
500 54] and requires some extent of human efforts in creating the data attributes, as it is still challenging
501 to automatically determine the optimal number/property of attributes without any labeled data and
502 task-specific information. Human-AI collaboration, while beneficial, may inevitably introduce certain
503 human biases into the model [43, 5]. We would like to clarify that in this work, the human raters are
504 blinded to the information for task-specific data during attribute selection. To further mitigate the
505 bias, it is crucial to involve multiple team raters, cross-validation, and additional review to achieve
506 a balanced and unbiased selection of attribute dimensions that contribute positively to the dataset’s
507 quality and diversity.

508 In addition, one issue with LLM-based training data generation is the phenomenon of *hallucination*,
509 wherein the model generates information that may sound plausible but is, in fact, not accurate or
510 grounded in reality. This can lead to the propagation of misinformation, particularly in sensitive areas
511 such as news generation where accuracy is of utmost importance. To alleviate this issue, it is possible
512 to leverage additional fact-checking mechanisms to cross-verify the generated text with a reliable
513 knowledge base or dataset. Furthermore, incorporating an additional layer of human review can also
514 help in mitigating hallucinations and ensuring the faithfulness of generated outputs.

515 **A.2 Societal impact**

516 The findings of this research have significant implications for the development of trustworthy machine
517 learning models. By addressing the limitations of simple class-conditional prompts and exploring
518 the potential of attributed prompts, the research aims to mitigate biases and enhance the diversity of
519 generated training data. This has a positive social impact as it contributes to the development of more
520 robust and fair machine learning models, reducing the potential for biased decision-making systems
521 and promoting inclusivity in various domains.

522 During the attribute selection process, we did not use any task-specific data to avoid data feature
523 leakage and minimize human bias. Other than this, we do not involve human subjects research and do
524 not contain any personally identifiable information. Possible misuse may lead to negative outcomes,
525 such as intentionally introducing biases to the generated dataset via attributes.

526 **A.3 Future work**

527 **Broader Coverage of Task and Modalities.** In future work, it would be valuable to extend
528 the investigation to other types of classification tasks beyond text classification. Exploring the
529 effectiveness of attributed prompts in image classification, audio classification, or other modalities
530 could provide insights into the broader applicability and effectiveness of AttrPrompt.

531 **Automatic Prompt Optimization.** In this work, we aim to demonstrate the usefulness of attributes
532 and why they are useful through the lens of diversity and bias, and we believe this opens the door
533 for future work to further explore the methodology for automatic attribute discovery and selection
534 for optimal performance, such as exploring automated or semi-automated methods for identifying
535 high-quality attribute dimensions and values. One promising direction is to incorporate a few labeled

536 examples as demonstrations for generating better training data. Under this scenario, leveraging
 537 Language Models can aid in automatically detecting diverse attributes, modeling the high-order
 538 interactions among attributes as well as estimating attribute value distributions for each class, which
 539 further reduce the human efforts for selecting key attributes.

540 B Workflow

The workflow of AttrPrompt is shown in Figure 7.

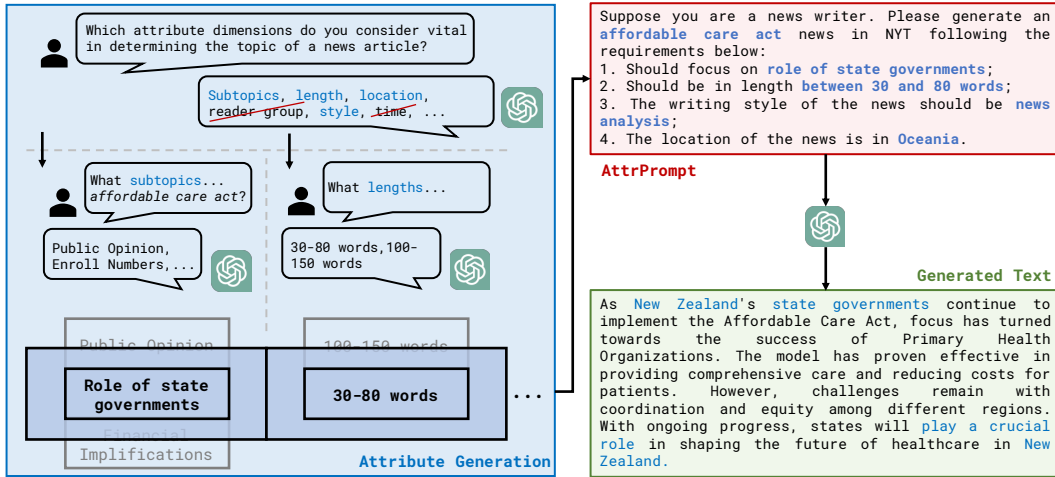


Figure 7: The overall workflow of AttrPrompt.

541

542 C Dataset Preprocessing

543 C.1 Multi-class Classification

544 For Reddit and StackOverflow, we select the classes with more than 65 examples from the original
 545 corpus as the target set of topics. For each dataset, we use 50 examples per class for the test set and no
 546 more than 10 examples for the validation set (10 for NYT/Amazon and 5 for Reddit/StackOverflow).
 547 The remaining data is used to compose the gold training set. It is worth noting that, some of the class
 548 names on Reddit may contain toxic information. To eliminate their effects, we filter our label names
 549 with Detoxify [19], a tool that leverages the multilingual XLM-RoBERTa [9] for toxic comment
 550 identification. We follow [14] to use a threshold of 0.1 to filter out potentially toxic topic classes.

551 C.2 Multi-label Classification

552 For the arXiv multi-label classification, we use the arXiv labels from the website⁶ as the label names.
 553 We select topics under the category of *Physics*, *Mathematics*, *Computer Science*, *Quantitative Biology*,
 554 and *Quantitative Finance*. Besides, we remove the topic of *General Physics*, *General Mathematics*,
 555 *General Literature*, *Other Computer Science*, *Other Quantitative Biology* as they only contain generic
 556 topics without concrete semantics. In total, there are 98 classes in general.

557 To generate examples using the SimPrompt approach, we follow a specific procedure. For each
 558 document, we begin by sampling from the distribution of label numbers, utilizing prior knowledge
 559 obtained from the original training set. This enables us to determine the precise number of labels,
 560 denoted as n , that should be associated with the document. Once we have determined n , we proceed

⁶<https://arxiv.org/>

561 to randomly select n classes from the original set of 98 labels. These chosen classes are then utilized
 562 as the input for prompting the LLM, forming a foundation for generating the examples.

563 To generate examples using the AttrPrompt approach, we employ a distinct methodology. For each
 564 document, we consider a set of attributes, namely subtopics, technique, writing style, and
 565 length. It is important to note that, unlike multi-class classification, where an additional step is
 566 typically taken to filter out irrelevant subtopics, our multi-label classification process follows a
 567 different approach. In our method, we introduce an additional step called *merging* to handle the
 568 different subtopics.

569 To determine the merging of subtopics, we utilize the fuzzy matching score⁷, which calculates the
 570 similarity between the subtopics. Setting the threshold at $\gamma = 90$, we proceed to merge subtopics
 571 whenever their similarity score exceeds this threshold. Consequently, specific subtopics may be
 572 associated with multiple classes, providing a more comprehensive representation of the document
 573 content. When generating text utilizing these subtopics, the corresponding labels are determined by
 574 the classes associated with the specific subtopics. This ensures that the generated examples accurately
 575 reflect the relevant classes based on the subtopic associations. The prompt format is deferred in
 576 section I.

577 D Implementation Details

578 D.1 Hardware information

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

581 D.2 Parameter configuration

582 We keep the parameter $\text{top_p} = 1.0$ and temperature $t = 1.0$ for calling ChatGPT APIs [38] for the
 583 training data generation part. For finetuning the classifier, we optimize the model using AdamW [30]
 584 with a linear warmup of the first 5% steps and linear learning rate decay. The maximum number of
 585 tokens per sequence is 128. Table 11 lists the hyperparameters used for AttrPrompt and SimPrompt.
 586 For the generated synthetic dataset, we stick to the strict zero-shot learning setting [31], train all
 587 the models for 6 epochs and use the model from the last epoch *without using the validation set* for
 588 evaluation. For the original clean dataset, we train all models for 6 epochs and use the model with the
 589 best performance on the validation set for evaluation.

Table 11: Hyperparameters for fine-tuning classifiers on different tasks.

Backbone	Learning Rate lr	Batch Size	Training Epochs E	Weight Decay	Warmup Ratio
BERT-base-uncased [12]	5e-5	32	6	1e-4	6%
TinyBERT [23]	1e-4	32	6	1e-4	6%
DistilBERT-base-uncased [46]	5e-5	32	6	1e-4	6%
DeBERTa-v3-base [21]	5e-5	32	6	1e-4	6%
DeBERTa-v3-large [21]	2e-5	32	6	1e-4	6%

Table 12: The reference of model checkpoints used in this study.

Model	Link
BERT-base-uncased [12]	https://huggingface.co/bert-base-uncased
TinyBERT [23]	https://huggingface.co/huawei-noah/TinyBERT_General_4L_312D
DistilBERT-base-uncased [46]	https://huggingface.co/distilbert-base-uncased
DeBERTa-v3-base [21]	https://huggingface.co/microsoft/deberta-v3-base
DeBERTa-v3-large [21]	https://huggingface.co/microsoft/deberta-v3-large
SentenceBERT [44]	https://huggingface.co/sentence-transformers/all-mpnet-base-v2
SimCSE [16]	https://huggingface.co/princeton-nlp/sup-simcse-bert-base-uncased
COCO-DR [56]	https://huggingface.co/OpenMatch/cocodr-base-msmarco

⁷<https://github.com/seatgeek/thefuzz>

590 **D.3 Model checkpoint**

591 All the checkpoint used in fine-tuning and linear probing experiments are listed in table 12.

592 **E Additional Experimental Results**

593 **E.1 Linear probing with generated datasets**

594 In this section, we evaluate the generated dataset under the linear probing setting, where a linear
 595 classifier is trained over the frozen pretrained features. To make a comprehensive evaluation, we
 596 employ four text embedding methods, namely BERT-CLS [12], SentenceBERT [44], SimCSE [16]
 597 and COCO-DR [56], and present the results in Table 13. We can see that AttrPrompt consistently
 598 outperforms SimPrompt, MetaPrompt and surpasses the Gold in the NYT dataset. This further
 599 indicates the datasets generated with AttrPrompt has higher quality than two baselines and this
 600 superiority is robust to the model choices.

Table 13: Performance of linear probing with different pre-trained encoders.

Sequence Encoder	Method	NYT		Amazon		Reddit		StackExchange	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
BERT-CLS [12]	Gold	65.39	70.26	66.73	68.76	55.73	58.00	17.17	17.68
	SimPrompt	60.69	62.26	37.95	38.07	27.41	27.60	12.27	13.71
	MetaPrompt	58.43	57.05	27.52	24.43	23.93	23.99	11.40	12.60
	AttrPrompt	62.38	64.17	48.18	49.46	45.03	45.95	17.33	17.64
SentenceBERT [44]	Gold	81.16	84.43	80.95	83.30	80.76	81.37	55.13	54.25
	SimPrompt	84.32	85.91	67.55	68.29	69.94	70.62	45.37	47.79
	MetaPrompt	87.65	87.11	67.16	65.78	79.26	69.15	45.71	47.54
	AttrPrompt	88.77	88.36	68.10	69.11	71.62	71.73	46.80	48.35
SimCSE [16]	Gold	78.02	80.17	73.20	75.13	73.75	74.26	45.26	41.86
	SimPrompt	78.93	78.86	52.40	52.83	54.78	53.86	30.98	32.92
	MetaPrompt	78.17	78.27	53.53	52.55	51.30	48.93	30.54	31.50
	AttrPrompt	81.41	81.04	60.66	61.32	59.22	59.15	30.31	31.43
COCO-DR [56]	Gold	80.66	82.60	76.13	78.34	79.63	80.04	61.02	63.77
	SimPrompt	84.37	84.17	55.66	55.28	63.66	64.71	43.22	41.50
	MetaPrompt	84.51	84.26	57.96	56.58	65.36	65.91	44.30	42.97
	AttrPrompt	86.74	86.43	60.47	62.03	66.73	66.71	45.44	43.76
Average	Gold	76.31	79.37	74.25	76.38	72.47	73.42	44.65	44.39
	SimPrompt	77.08	77.80	53.39	53.62	53.95	54.20	32.96	33.98
	MetaPrompt	77.19	76.67	51.54	49.84	54.96	52.00	32.99	33.65
	AttrPrompt	79.83	80.00	59.35	60.48	60.65	60.89	34.97	35.30

601 **E.2 Additional results on attribute selection**

602 Table 14 shows the result of AttrPrompt without human selection of attributes on two datasets.
 603 While a slight performance drop is observed, the result of AttrPrompt without attribute selection still
 604 significantly outperform the SimPrompt. This demonstrates the robustness of AttrPrompt performance
 605 to potentially irrelevant attributes.

	NYT	Amazon
GPT-3.5 Zero-shot	69.84	54.56
SimPrompt	76.34	56.96
AttrPrompt	82.26	65.87
AttrPrompt w/o selection	81.08	63.76

Table 14: The Performance of AttrPrompt with/without attribute selection.

606 **E.3 Additional results on data/budget efficiency**

607 In Section 5.3, we discuss the data/budget efficiency of AttrPrompt and SimPrompt on NYT and
 608 Amazon datasets; here, we provide additional results on the remaining datasets in Figure 8. We
 609 draw conclusions similar to what is stated in Section 5.3 that AttrPrompt exhibits better data/budget
 610 efficiency than SimPrompt.

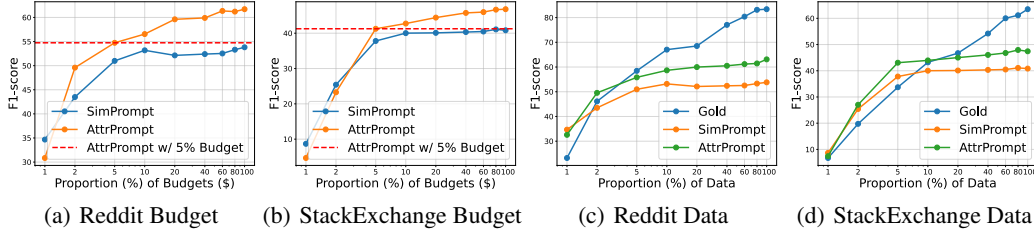


Figure 8: The comparisons on budget efficiency and data efficiency on Reddit and StackExchange.

611 **E.4 Different temperature parameters for ChatGPT**

612 Temperature (t) is one crucial hyperparameter of LLMs that controls the diversity of the generated
 613 text [20, 10], while the studied attributed prompts are also for diversifying the generated data. We are
 614 then curious about the effectiveness of the temperature and how it compares to the AttrPrompt. We
 615 study different values of the temperature using the NYT dataset and present the results in Table 15.
 616 From the results, we can see that compared with the temperature, AttrPrompt brings more significant
 617 performance gain, demonstrating its superiority over temperature tuning.

Table 15: 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

618 **E.5 The performance with respect to long-tail classes**

619 As we have seen in Table 2, the original training sets of the involved datasets have severe long-tail
 620 classes issue since the imbalance ratio is high, yet the generated dataset are class-balanced, we are then
 621 curious how the class balance in the generated dataset benefits the model performance on long-tail
 622 classes. We take the NYT dataset as an example and plot the per-class F1 score of Gold, SimPrompt,
 623 and AttrPrompt in Figure 9, where the x-axis is classes sorted by their number of data in the Gold
 624 dataset in descending order. From the figure, we can see that out of 26 classes, AttrPrompt renders the
 625 best per-class F1 score on 10 classes, which is 13 for Gold and 3 for SimPrompt. Notably, for classes
 626 with few examples in the Gold set (the rightmost 4 classes in the figure), AttrPrompt is better than the
 627 Gold and SimPrompt, especially for the class “*abortion*” with the fewest examples. This suggests
 628 a data-centric way to handle the long-tail class issue in topic classification: one may use LLMs to
 629 generate class-balanced training set or augment the existing training set with the LLM-generated data
 630 such that the augmented dataset is class-balanced, the in-depth study of which is left as future work.

631 **E.6 Experiment results on other datasets**

632 To provide a comprehensive evaluation of AttrPrompt, we conducted additional empirical studies
 633 using datasets commonly employed in prior works, which typically exhibit *low cardinality*. Consistent
 634 with [55, 54, 57], we focused on four specific tasks: IMDB, SST-2, Yelp, and AG News. The IMDB,
 635 SST-2, and Yelp datasets involve binary sentiment classification, while the AG News dataset entails
 636 news classification with only four classes. This selection allows for diverse evaluation across different
 637 classification scenarios.

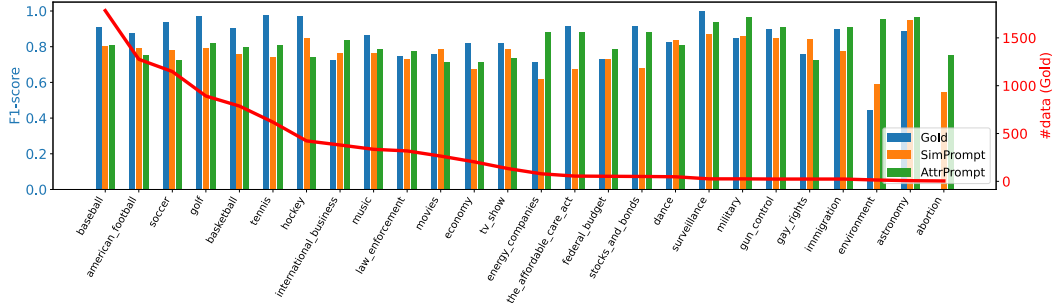


Figure 9: Per-class F1-score of the NYT dataset.

Table 16: Results with recent baselines using DistilBERT [46]. *: Results are copied from the previous papers [55, 54, 15, 57].

Method/Dataset	IMDB	SST-2	Yelp	AG News	Avg.
<i>Simple fine-tuning with standard cross-entropy loss</i>					
SimPrompt	85.64	79.38	88.39	74.30	81.93
AttrPrompt	85.92	83.64	90.60	82.74	85.73
<i>Other results with dedicated training techniques (for reference only)</i>					
Zero-shot Prompting*	77.31	82.63	78.10	73.20	77.81
ZeroGen*	82.77	80.41	87.84	76.48	81.88
SuperGen*	85.05	84.28	89.88	77.40	84.15
ProGen*	87.20	84.12	89.39	80.81	85.38
SunGen*	89.45	89.38	89.19	80.49	87.12
ReGen*	85.32	87.84	89.00	80.18	85.58

638 To ensure a fair comparison, we utilized a smaller pre-trained language model (PLM) called Distill-
639 BERT [46] as the classifier, in line with [55, 54, 15]. For the purpose of fine-tuning the classifier, we
640 generated a total of 6000 examples, evenly distributed among the classes. For fine-tuning, we use a
641 learning rate of 2×10^{-5} with 5 epochs and use the model after the final epoch for evaluation.

642 The results of our evaluation are presented in table 16. Overall, we observe that AttrPrompt demon-
643 strates competitive performance across all four tasks. It outperforms our direct baseline (SimPrompt)
644 by an impressive margin of 3.8%. Moreover, when compared to other baselines incorporating
645 noisy-aware learning techniques, AttrPrompt still achieves promising performance, despite using
646 the standard cross-entropy loss. This suggests that in addition to leveraging noisy-robust learning
647 objectives, enhancing the diversity of input prompts serves as an effective alternative to improving
648 downstream performance. It is interesting yet important to combine these different learning paradigms
649 together to further boost the performance.

650 F Regional Bias with Manual Annotations

651 In the main body of the paper, we study the regional bias in the NYT dataset via an attribute classifier
652 trained with the location attribute of the prompt associated with each generated data. Here, we
653 manually annotate location for 100 samples from each of the Gold, SimPrompt, and AttrPrompt
654 dataset, and compare manual annotations with the prediction of the attribute classifier in Table 17
655 (the first three rows). Note that we include “*unkown*” as an option in manual annotation to absorb text
656 without clear location specification. From the results, we can see that the attribute classifier largely
657 aligns with manual annotations, since the accuracy is greater than 0.95 on the Gold and AttrPrompt
658 and 0.82 on SimPrompt; and the result on SimPrompt is lower since SimPrompt tend to generate
659 more unknown location data.

660 With manual annotations, we also check how well ChatGPT can follow the required location attribute
661 in the prompts. We compute the accuracy between the location attribute in the prompt associated with

662 each data and that annotated by human on the 100 AttrPrompt samples (the last row of Table 17),
 663 which is 0.96. This means that ChatGPT is able to faithfully generate news with desired location
 664 specifications.

Table 17: Accuracies based on manual annotation of the location attribute of the NYT dataset.

Comparison	Data Source	Accuracy
Attribute Classifier v.s. Human	Gold	0.96
	SimPrompt	0.82
	AttrPrompt	0.95
Attribute in Prompt v.s. Human	AttrPrompt	0.96

665 We also plot the pie charts using manual annotations of location in Figure 10, and obtain similar
 666 findings as in Section 4.2, *i.e.*, the data in Gold and SimPrompt are largely biased towards “North
 667 America”, while that of AttrPrompt is relatively balanced.

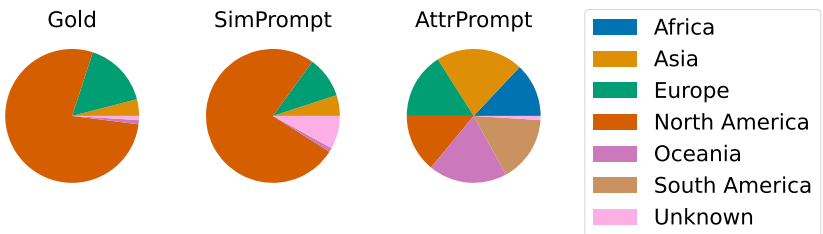


Figure 10: Pie charts of the “location” attribute based on manual annotations.

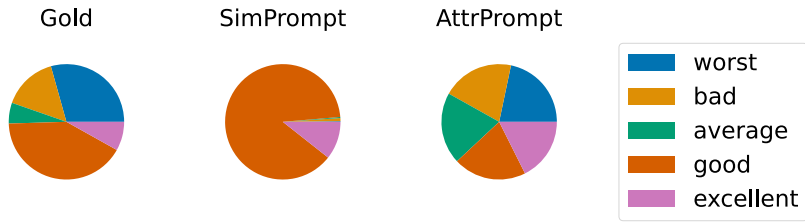
668 G Additional Bias Analysis

669 G.1 Biases other than regional bias

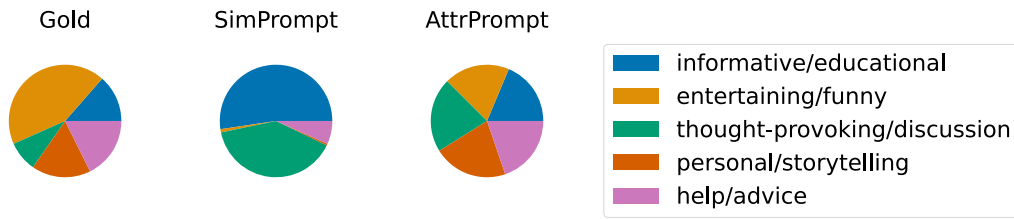
670 In addition to the regional bias we study in the main body of the paper, here we investigate bias in
 671 other datasets (Figure 11). From the results, we can see that the dataset generated with Gold and
 672 SimPrompt exhibit different level of biases, which shows that in general the biases exists in both
 673 generated and real dataset, while dataset generated with AttrPrompt has relatively balanced attribute
 674 distribution.

675 G.2 Bias analysis in AttrPrompt

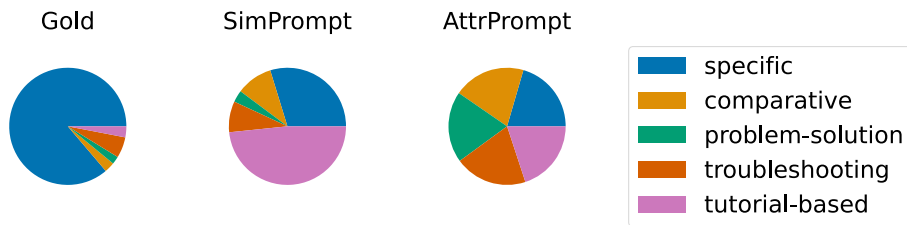
676 Here, we compare the distribution of predicted “location” in the dataset generated with AttrPrompt
 677 to that with the Gold and SimPrompt in Figure 12. We can see that AttrPrompt renders a relatively
 678 balanced “location” distribution.



(a) Amazon-Experience.



(b) Reddit-Style.



(c) StackExchange-Style.

Figure 11: Pie charts of predicted attribute distributions on different datasets.

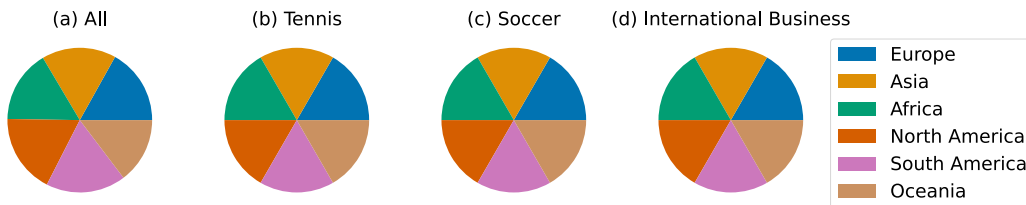


Figure 12: Pie charts of the distributions of “location” predicted by an attribute classifier for the NYT AttrPrompt dataset. (a) is “location” distribution over the whole dataset, while others are specific classes.

679 H The Prompt Format for LLM-Generated Attributes

680 H.1 Prompts for NYT Dataset

681 H.1.1 Key Attributes

682 **Prompts to ChatGPT.** What do you think are important attributes to generate some diverse news
683 from NYT under a specific topic, such as Football, Federal Budget. Examples: subtopics, writing
684 style...

685 **Response from ChatGPT.** Important attributes of news from the New York Times under a specific
686 topic could include:

687 Subtopics: exploring different angles and subtopics within the main topic can lead to a wider range
688 of news stories.

689 Writing style: varying the writing style can help to make the news stories more interesting and
690 engaging for readers. This can include using different tones, perspectives, and sentence structures.

691 Sources: using a variety of sources, such as experts, eyewitnesses, and affected individuals, can
692 provide different perspectives and add diversity to the news stories.

693 Geographical locations: covering the topic from different geographical locations can help to show
694 how it affects different communities and provide a global perspective.

695 Timeliness: covering current events and breaking news related to the topic can add diversity and
696 relevance to the news stories.

697 Multimedia: using a range of multimedia elements, such as images, videos, and infographics, can
698 help to add variety and appeal to the news stories.

699 Length: varying the length of the news stories, from short and concise to longer and more in-depth,
700 can help to provide a wider range of information and cater to different readers' preferences.

701 Selected Attributes

- 702 1. Subtopics
- 703 2. Writing Styles
- 704 3. Geographical locations (determined by six continents by default)
- 705 4. Length (determined by {30-80} and {100-150} by default)

706 H.1.2 Attributes Details

707 **Prompts to ChatGPT for Similar Classes.** List 5 similar classes for {class-name} news on NYT.
708 The set of classes is listed as: {[a list of class-names]}.

709 **Prompts to ChatGPT for Subtopics.** List 10 diverse subtopics for {class_name} news on NYT.

710 **Response from ChatGPT.** See section I for details.

711 H.1.3 Writing Styles

712 **Prompts to ChatGPT.** List 5 different writing styles of NYT news.

713 Response from ChatGPT.

- 714 1. Investigative reporting
- 715 2. Opinion/editorial
- 716 3. Human interest/profiles

717 4. Long-form narrative

718 5. Breaking news

719 H.2 Prompts for Amazon Dataset

720 H.2.1 Key Attributes

721 **Prompts to ChatGPT.** What do you think are important attributes to generate some diverse
722 scientific paper abstracts under a specific topic, like computer vision, machine learning, etc. Examples:
723 subtopics, keywords, writing style...

724 **Response from ChatGPT.** When writing a review for a product, here are five important attributes
725 to consider:

726 **Product Type:** Clearly mention the type of product you are reviewing, such as a smartphone, laptop,
727 or fitness tracker. This helps readers understand the category and purpose of the product.

728 **Brand:** Specify the brand of the product as it often influences quality, reputation, and customer
729 support. Discuss the brand's overall credibility and whether it aligns with your expectations.

730 **User Experience:** Evaluate the overall user experience of the product. Discuss its speed, accuracy,
731 reliability, and efficiency in performing its intended tasks. Highlight any exceptional or lacking
732 performance aspects.

733 **Quality and Durability:** Assess the quality of the product, including the materials used, construction,
734 and overall durability. Discuss whether it feels well-made, solid, and likely to withstand regular use
735 over time.

736 **Features and Functionality:** Describe the specific features and functions of the product. Highlight
737 any unique or standout features that enhance its usability or set it apart from similar products in the
738 market.

739 H.2.2 Selected Attributes

740 1. Product Name

741 2. Brands

742 3. User Experience

743 4. Writing Styles

744 5. Length

745 H.2.3 Attributes Details

746 **Prompts to ChatGPT for Similar Classes.** List 5 similar classes for `{product-class}` products
747 on Amazon. The set of classes is listed as: `{[a list of class-names]}`.

748 **Prompts to ChatGPT for Product Name/Brands.** List 20 diverse product names/brands for
749 `{product-class}` on Amazon.

750 **Response from ChatGPT.** See section [I.2](#) for details.

751 H.2.4 Usage Experience

752 **Prompts to ChatGPT.** List 5 different usage experiences of Amazon reviews.

753 **Response from ChatGPT.**

754 1. Worst: A highly unsatisfactory or problematic experience. Users may encounter multiple
755 issues, such as usability problems, errors, and frustrations.

- 756 2. Bad: A below-average experience. Users may encounter some difficulties or frustrations
 757 while interacting with the product or service, but the overall experience falls short of meeting
 758 their expectations.
- 759 3. Average: A neutral or middling experience. Users may not encounter major issues, but the
 760 experience lacks any exceptional aspects or standout features.
- 761 4. Good: A positive and satisfactory experience. Users have a smooth and enjoyable experience
 762 with minimal issues or obstacles.
- 763 5. Excellent: The highest level of satisfaction. Users have a seamless, delightful, and memo-
 764 rable experience.

765 **H.2.5 Writing Styles**

766 **Prompts to ChatGPT.** List 5 different writing styles of Amazon reviews.

767 **Response from ChatGPT.**

- 768 1. Detailed Review: A detailed review provides a comprehensive analysis of the product,
 769 covering various aspects such as features, performance, durability, and user experience.
- 770 2. Comparative Review: A comparative review compares the product with similar options
 771 available on Amazon.
- 772 3. Pros and Cons Review: A pros and cons review presents a balanced assessment of the
 773 product by listing its positive and negative aspects.
- 774 4. Personal Experience Review: A personal experience review shares the reviewer’s firsthand
 775 experience with the product.
- 776 5. Recommendation Review: A recommendation review offers a clear recommendation or
 777 endorsement of the product.

778 **H.2.6 Prompt Template**

See table 18 for details.

Table 18: Prompt template for the Amazon Product Review dataset.

Method	Prompt
SimPrompt	Suppose you are a review writer. Please write a review for {product-class} in Amazon.
AttrPrompt	Suppose you are a review writer. Please write a review for {product-class} product in Amazon following the requirements below: 1. The review should be about the product of {subtopic}; 2. The brand for the product should be {brand}; 3. Should be in length between {length:min-words} and {length:max-words} words; 4. Should describe the usage experience {usage-experience}; 5. The writing style of the review should be {style}; 6. the review must be relevant to {product-class} and irrelevant to: {similar-class}.

779

780 **H.3 Prompts for Reddit Dataset**

781 **H.3.1 Key Attributes**

782 **Prompts to ChatGPT.** List 5 important attributes to generate some diverse reddit posts under a
 783 specific topic, like build_a_pc, premed, etc. Examples: subtopics, keywords, writing style...

784 **Response from ChatGPT.** Informative Guides and Tutorials: Create posts that provide step-
 785 by-step guides, tutorials, or informative content. These can include detailed instructions, tips, or
 786 recommendations to help readers in their journey. For example, in "build_a_pc," you could create a
 787 guide on assembling a gaming PC or optimizing system performance.

788 Writing Style: Utilize different writing styles to create diverse posts. For instance, you can write
789 informative guides or tutorials, opinion pieces, comparison reviews, troubleshooting tips, or personal
790 stories related to the topic.

791 Discussion Starters: Pose thought-provoking questions or statements to initiate discussions and
792 engage the Reddit community. These can be open-ended queries, asking for opinions or experiences,
793 or presenting a controversial viewpoint to encourage debate and dialogue.

794 Personal Experiences: Share personal stories or experiences related to the topic. This allows you to
795 connect with the community on a more personal level and provide insights based on your firsthand
796 experiences. It can include successes, challenges, lessons learned, or interesting anecdotes.

797 Resources and Recommendations: Provide valuable resources such as articles, websites, tools, or
798 recommended products/services related to the topic. These resources can help the community access
799 additional information, useful tools, or trusted recommendations, enhancing their understanding and
800 engagement with the subject.

801 H.3.2 Selected Attributes

- 802 1. Personal Experiences
- 803 2. Resources
- 804 3. Writing Styles
- 805 4. Length

806 H.3.3 Attributes Details

807 **Prompts to ChatGPT for Similar Classes.** List 5 similar classes for `{post-class}` products on
808 Reddit. The set of classes is listed as: `{[a list of class-names]}`.

809 **Prompts to ChatGPT for Personal Experiences/Resources.** List 20 diverse Personal Experi-
810 ences/Resources for posts about `{post-class}` in Reddit.

811 **Response from ChatGPT.** See section I.3 for details.

812 H.3.4 Writing Styles

813 **Prompts to ChatGPT.** List 5 different writing styles of Reddit Posts.

814 **Response from ChatGPT.**

- 815 1. Informative/Educational: Create posts that aim to provide valuable information or educate
816 the community on a particular topic. Share informative articles, research findings, guides,
817 tutorials, or explain complex concepts in a simplified manner.
- 818 2. Entertaining/Funny: Craft posts that are meant to entertain or bring humor to the community.
819 Share funny stories, jokes, memes, amusing videos, or any form of content that can elicit
820 laughter and enjoyment.
- 821 3. Thought-Provoking/Discussion: Initiate posts that spark thoughtful discussions and invite
822 users to share their opinions or experiences. Pose open-ended questions, present philo-
823 sophical dilemmas, discuss current events, or explore societal issues to foster engaging
824 conversations.
- 825 4. Personal/Storytelling: Share personal experiences, anecdotes, or narratives that others can
826 relate to. Tell compelling stories, share achievements, challenges, or life lessons that can
827 inspire or resonate with the community.
- 828 5. Help/Advice: Create posts where you seek help or offer advice to the community. Ask for
829 guidance on a specific problem, seek recommendations, provide support, or share expertise
830 in a particular field.

Table 19: Prompt template for the Reddit dataset.

Method	Prompt
SimPrompt	Suppose you are a writer for posts in Reddit. Please give a post for <code>{post-class}</code> in Reddit.
AttrPrompt	Suppose you are a review writer. Please give a post in Reddit on <code>{post-class}</code> community following the requirements below: 1. The post should focus on <code>{experience}</code> ; 2. Should be in length between <code>{length:min-words}</code> and <code>{length:max-words}</code> words; 3. The writing style of the post should be <code>{style}</code> ; 4. Should mention the resource of <code>{resource}</code> ; 5. the post must be relevant to <code>{post-class}</code> and irrelevant to: <code>{similar-class}</code> .

831 **H.3.5 Prompt Template**

832 See table 19 for details.

833 **H.4 Prompts for StackExchange Dataset**

834 **H.4.1 Key Attributes**

835 **Prompts to ChatGPT.** What do you think are important attributes to generate diverse posts under a
836 specific topic on the stackexchange forum, like 3D meshes, server, etc. For example, one important
837 attribute could be writing style...

838 **Response from ChatGPT.** When generating diverse posts under a specific topic on the Stack
839 Exchange forum, several important attributes can contribute to the variety and richness of the
840 discussions:

841 Depth of Knowledge: Participants with different levels of expertise can offer insights and perspectives
842 from beginner to advanced, allowing for a broader range of questions and answers.

843 Diverse Perspectives: Encouraging participants from various backgrounds, industries, and experiences
844 can bring fresh viewpoints and approaches to the topic.

845 Objectives: Encouraging participants to state their objectives or goals in their posts can help generate
846 diverse content. Some posts may aim to seek advice, share experiences, provide step-by-step guides,
847 analyze specific problems, or discuss broader concepts.

848 Problem Scenarios: Encourage users to discuss real-world problem scenarios related to the topic.
849 This helps generate diverse posts by exploring different use cases, challenges, and approaches to
850 problem-solving.

851 Problem Solving: Promoting problem-solving skills and encouraging users to share their challenges,
852 roadblocks, and solutions can foster a collaborative and dynamic environment.

853 Clear Communication: Emphasizing clear and concise communication helps to ensure that questions
854 are well-articulated and that answers are easily understood by a wide range of users.

855 **H.4.2 Selected Attributes**

- 856 1. Problem Scenarios
- 857 2. Depth of Knowledge
- 858 3. Writing Styles
- 859 4. Length

860 **Prompts to ChatGPT for Problem Scenarios.** List 20 diverse Problem Scenarios for posts about
861 `{post-class}` in StackExchange.

862 **Response from ChatGPT.** See section I.4 for details.

863 **H.4.3 Writing Styles**

864 **Prompts to ChatGPT.** List 5 different writing styles of StackExchange Posts.

Table 20: Prompt template for the StackExchange dataset.

Method	Prompt
SimPrompt	Suppose you are a writer for posts in StackExchange. Please give an example of a question post in StackExchange about <code>{post-class}</code> .
AttrPrompt	<p>Please give an example of a question post in StackExchange about <code>{post-class}</code> following the requirements below:</p> <ol style="list-style-type: none"> 1. Should focus on the scenario of <code>{scenario}</code>; 2. Should be in length between <code>{length:min-words}</code> and <code>{length:max-words}</code> words; 3. The writing style of the question should be <code>{style}</code>; 4. The question should be in <code>{depth}</code>; 5. The review must be relevant to <code>{post-class}</code> and irrelevant to: <code>{similar-class}</code>.

865 **Response from ChatGPT.**

- 866 1. Specific: Be specific in your question by including relevant details, such as error messages,
867 code snippets, or specific scenarios, to help the community provide accurate and targeted
868 answers.
- 869 2. Comparative: Compare different approaches, technologies, or methodologies, and ask for
870 insights or opinions from the community on their strengths, weaknesses, and suitability for
871 your specific situation.
- 872 3. Problem-Solution: Present a problem or challenge you are encountering and outline any
873 attempted solutions, seeking feedback or alternative approaches from the community to help
874 you find an optimal solution.
- 875 4. Troubleshooting: Describe a specific issue or error you are encountering and provide any
876 relevant error messages or logs, seeking assistance in identifying the cause and potential
877 solutions.
- 878 5. Tutorial-Based: Ask for guidance or step-by-step instructions on how to achieve a specific
879 task or goal, seeking expertise and advice from the community.

880 **H.4.4 Depth of Knowledge**

881 **Prompts to ChatGPT.** List 5 different levels of depth of knowledge for StackExchange Posts.

882 **Response from ChatGPT.**

- 883 1. Beginner Level: Questions that address basic concepts and seek introductory explanations
884 or guidance. These questions are suitable for individuals who are new to the topic.
- 885 2. Intermediate Level: Questions that require a moderate understanding of the topic and seek
886 more specific information or solutions to intermediate-level problems.
- 887 3. Advanced Level: Questions that delve into complex concepts or advanced techniques,
888 requiring a deeper understanding of the topic and seeking advanced solutions or insights.
- 889 4. Expert Level: Questions that involve specialized or advanced knowledge within a specific
890 field, targeting experts or professionals seeking detailed and intricate solutions.

891 **H.4.5 Prompt Template**

892 See table 21 for details.

893 **H.5 Prompts for arXiv Dataset**

894 **H.5.1 Key Attributes**

895 **Prompts to ChatGPT.** What do you think are important attributes to generate diverse papers under a
896 specific topic on arXiv. For example, one important attribute could be writing style...

897 **Response from ChatGPT.** Generating diverse papers under a specific topic on platforms like arXiv
898 involves considering various important attributes. Here are a few attributes that can contribute to
899 generating diverse papers:

Table 21: Prompt template for the arXiv multi-label dataset.

Method	Prompt
SimPrompt	Please give a synthetic paper abstract about <code>{paper-classes}</code> .
AttrPrompt	Please write an abstract of a <code>{paper-classes}</code> paper in arXiv following the requirements below: 1. Should focus on the <code>{subtopic}</code> ; 2. Should be in length between <code>{length:min-words}</code> and <code>{length:max-words}</code> words; 3. The paper should use the technique relevant to <code>{technique}</code> ; 4. The style of the paper should be in <code>{style}</code> ;

900 Writing Style: Encouraging papers with different writing styles, such as technical, expository,
901 theoretical, or empirical, can bring diversity to the presentation and appeal to a wider range of readers.

902 Subtopics: Promoting papers that explore different subtopics within the broader topic can provide
903 comprehensive coverage and delve into specific areas of interest.

904 Techniques: Encouraging papers that employ different research methodologies, such as experimental,
905 computational, or analytical, can bring diverse approaches to studying the topic.

906 Data Sources: Promoting papers that utilize diverse data sources, such as surveys, simulations,
907 real-world datasets, or case studies, can offer different perspectives and insights into the topic.

908 Interdisciplinary Perspectives: Encouraging papers that incorporate interdisciplinary perspectives,
909 drawing insights from multiple fields or combining methodologies from different disciplines, can
910 contribute to a richer understanding of the topic.

911 H.5.2 Selected Attributes

912 1. Subtopics

913 2. Techniques

914 3. Writing Styles

915 4. Length

916 **Prompts to ChatGPT for Subtopics/Techniques.** List 20 diverse subtopics/techniques for paper
917 about `{paper-class}` in arXiv.

918 **Response from ChatGPT.** See section I.5 for details.

919 H.5.3 Writing Styles

920 **Prompts to ChatGPT.** List 5 different writing styles of arXiv papers.

921 **Response from ChatGPT.**

922 1. Experimental: This style focuses on presenting empirical research with a strong emphasis
923 on experimental design, data collection, and statistical analysis.

924 2. Theoretical: This style emphasizes the development of new theoretical frameworks or
925 conceptual models.

926 3. Review: This style involves synthesizing and analyzing existing literature on a specific topic.

927 4. Case Study: This style presents an in-depth examination of a specific case or situation,
928 aiming to provide insights into complex phenomena or phenomena that are context-specific.

929 H.5.4 Prompt Template

930 See table 21 for details.

931 **H.6 Prompts for AG News Dataset**

932 The prompts are the same as the NYT dataset. The only difference is the subtopics for different labels.
933 The subtopic information is listed in [I.7](#) for details.

934 **H.7 Prompts for Yelp/SST-2/IMDB Dataset**

935 **H.7.1 Key Attributes**

936 **Prompts to ChatGPT.** What do you think are important attributes to generate some diverse reviews
937 with sentiments? E.g. writing styles

938 **Response from ChatGPT.** Writing Styles: Utilizing different writing styles such as descriptive,
939 narrative, persuasive, or humorous can create a diverse range of tones and emotions in the reviews.

940 Language and Tone: Varying the language and tone used in the reviews can evoke different sentiments.
941 This can include formal or casual language, positive or negative tone, or even incorporating elements
942 of storytelling.

943 Personal Experience: Sharing personal experiences and anecdotes in the reviews can add authenticity
944 and emotional depth, allowing readers to connect with the sentiments expressed.

945 Descriptive Details: Including specific and vivid descriptions of the ambiance, decor, flavors, smells,
946 or textures of the food can create a sensory experience that evokes various sentiments.

947 **H.7.2 Selected Attributes**

- 948 1. Personal Experience
- 949 2. Descriptive Details
- 950 3. Writing Styles
- 951 4. Length

952 **H.7.3 Personal Experiences/Descriptive Details**

953 **Prompts to ChatGPT.** Details are listed in Appendix [I.7](#) and [I.8](#).

954 **H.7.4 Writing Styles**

955 **Prompts to ChatGPT.** List 5 different writing styles of Restaurant Reviews (for Yelp).

956 **Response from ChatGPT.**

- 957 1. Descriptive Style: Focuses on providing detailed descriptions of the restaurant's ambiance,
958 decor, and overall atmosphere. It aims to paint a vivid picture of the dining experience.
- 959 2. Personal Narrative Style: Incorporates personal anecdotes and experiences from the visit to
960 the restaurant. It adds a personal touch to the review and allows the reader to connect with
961 the reviewer's perspective.
- 962 3. Analysis Style: Takes a more analytical approach by evaluating various aspects of the
963 restaurant, such as food quality, service, cleanliness, and value for money.
- 964 4. Comparative Style: Compares the restaurant to similar establishments in terms of cuisine,
965 price range, service quality, or other relevant factors.
- 966 5. Gastronomic Style: Focuses primarily on the food and beverage offerings of the restaurant.

967 **Prompts to ChatGPT.** List 5 different writing styles of Movie Reviews (for IMDB/SST-2).

Table 22: Prompt template for the Yelp dataset.

Method	Prompt
SimPrompt	Suppose you are a review writer. Please give an example of a {sentiment} review in Yelp.
AttrPrompt	Please write a {sentiment} review for a restaurant, following the requirements below: 1. The overall review should be {sentiment}; 2. It should be a {descriptive-details} restaurant ; 3. Should focus on {personal-experience}; 4. Should be in length between {length:min-words} and {length:max-words} words; 5. The writing style of the review should be {style};

Table 23: Prompt template for the SST-2/IMDB dataset.

Method	Prompt
SimPrompt	Suppose you are a review writer. Please give an example of a {sentiment} review for a movie.
AttrPrompt	Please write a {sentiment} review for a movie, following the requirements below: 1. The overall review should be {sentiment}; 2. It should be a {descriptive-details} movie ; 3. Should focus on {personal-experience}; 4. Should be in length between {length:min-words} and {length:max-words} words; 5. The writing style of the review should be {style};

968 **Response from ChatGPT.**

- 969 1. Descriptive: Focuses on providing a detailed and vivid description of the movie and its
 970 elements.
- 971 2. Analytical: Analyzes and evaluates various aspects of the movie, such as the plot, acting,
 972 cinematography, and direction.
- 973 3. Persuasive: Aims to convince the reader to watch or avoid the movie by presenting strong
 974 arguments and opinions.
- 975 4. Comparative: Compares the movie to other films within the same genre or by the same
 976 director.

977 **H.7.5 Prompt Template**

978 See table 22 and 23 for details.

979 **I Details for LLM-Generated Attributes**

980 **I.1 NYT Attributes**

981 **I.1.1 Subtopics**

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

- 984 • astronomy:
- 985 – Discoveries of exoplanets
 - 986 – Black holes and their role in shaping galaxies
 - 987 – The search for extraterrestrial life
 - 988 – Gravitational waves and the study of the universe’s origins
 - 989 – The use of telescopes to explore the universe
 - 990 – The mysteries of dark matter and dark energy
 - 991 – Solar flares and their impact on Earth
 - 992 – The history of the universe and its evolution over time
 - 993 – Exploring the possibility of space tourism
 - 994 – The exploration of our neighboring planets, such as Mars and Venus.

- 995 • baseball:
 - 996 – Recent controversy surrounding sign-stealing scandal in MLB
 - 997 – Breakdown of top prospects in minor league baseball
 - 998 – Analysis of new rule changes for upcoming baseball season
 - 999 – Coverage of recent World Series champions and their success
 - 1000 – In-depth profile of influential baseball figures, such as managers or players
 - 1001 – Updates on retired players and their post-baseball careers
 - 1002 – Highlighting standout performances by individual players or teams in recent games
 - 1003 – Coverage of international baseball leagues and their top players
- 1004 • economy:
 - 1005 – Job market and employment rates
 - 1006 – Interest rates and monetary policy
 - 1007 – Inflation and deflation
 - 1008 – Economic growth and GDP
 - 1009 – Consumer spending and retail sales
 - 1010 – Income inequality and poverty
 - 1011 – GDP growth and contraction
 - 1012 – Labor market trends
 - 1013 – Economic impacts of natural disasters and pandemics
 - 1014 – Housing market and real estate
- 1015 • federal_budget:
 - 1016 – Deficit reduction strategies
 - 1017 – Government spending priorities
 - 1018 – Tax reform proposals
 - 1019 – Agency budget cuts and restructuring
 - 1020 – Budget negotiations and debates
 - 1021 – National debt projections
 - 1022 – Fiscal policy decisions
 - 1023 – Congressional budget proposals
 - 1024 – Infrastructure spending plans
 - 1025 – Public opinion on federal budget issues
- 1026 • movies:
 - 1027 – Box office records and trends for Hollywood blockbusters
 - 1028 – Pioneering techniques in film-making and special effects
 - 1029 – Representation and diversity in casting and storytelling
 - 1030 – Reviews and analysis of highly-anticipated new releases
 - 1031 – The impact of streaming services on movie distribution and consumption
 - 1032 – The intersection of politics and film, from socially-conscious storytelling to politically-charged controversies
 - 1033 – Profiles of notable actors, directors, and producers shaping the industry
 - 1034 – The changing landscape of film festivals and awards season
 - 1035 – Spotlight on independent and international cinema
 - 1036 – The legacy of classic films and their enduring cultural impact.
 - 1037

1038 **I.2 Amazon Attributes**

1039 **I.2.1 Product Brands**

1040 We randomly select 5 categories in Amazon dataset and display the corresponding product brand
1041 attributes for each category:

- 1042 • camera_photo.:
 - 1043 – Canon
 - 1044 – Nikon
 - 1045 – Leica
 - 1046 – Hasselblad
 - 1047 – Fujifilm
 - 1048 – Lumix (Panasonic)
 - 1049 – GoPro
 - 1050 – Polaroid
 - 1051 – Yashica
 - 1052 – Mamiya
- 1053 • jewelry_and_watches.:
 - 1054 – Rolex
 - 1055 – Cartier
 - 1056 – Tiffany & Co.
 - 1057 – Bulgari
 - 1058 – Omega
 - 1059 – Patek Philippe
 - 1060 – Swarovski
 - 1061 – Gucci
 - 1062 – Michael Kors
 - 1063 – Pandora
- 1064 • magazines.:
 - 1065 – Rolling Stone
 - 1066 – Vogue
 - 1067 – National Geographic
 - 1068 – The New Yorker
 - 1069 – GQ
 - 1070 – Cosmopolitan
 - 1071 – People
 - 1072 – Time
 - 1073 – Sports Illustrated
 - 1074 – Forbes
- 1075 • health_and_personal_care.:
 - 1076 – Johnson & Johnson
 - 1077 – Dove
 - 1078 – Colgate
 - 1079 – Head & Shoulders
 - 1080 – Merck
 - 1081 – Pfizer
 - 1082 – Merck

- 1083 – Abbott Laboratories
- 1084 – GlaxoSmithKline
- 1085 • toys_games.:
- 1086 – Mattel
- 1087 – Fisher-Price
- 1088 – Hasbro
- 1089 – Lego
- 1090 – Nerf
- 1091 – Barbie
- 1092 – Hot Wheels
- 1093 – Playmobil
- 1094 – MGA Entertainment
- 1095 – Paw Patrol

1096 **I.2.2 Product Names**

1097 We randomly select 5 categories in Amazon dataset and display the corresponding product name
1098 attributes for each category:

- 1099 • sports_outdoors.:
- 1100 – Trekking poles
- 1101 – Kayak
- 1102 – Stand-up paddleboard
- 1103 – Treadmill
- 1104 – Bike
- 1105 – Yoga mat
- 1106 – Weightlifting gloves
- 1107 – Athletic training cones
- 1108 – Ab wheel
- 1109 – Resistance bands
- 1110 – Jump rope
- 1111 – Rollerskates
- 1112 – Boxing gloves
- 1113 – Basketball
- 1114 – Football
- 1115 – Golf clubs
- 1116 – Tennis racquet
- 1117 • camera_photo.:
- 1118 – SnapShooter
- 1119 – FocusReady
- 1120 – ClickCapture
- 1121 – MemoriesMatter
- 1122 – FlashFinder
- 1123 – PicturePerfect
- 1124 – ShotSnap
- 1125 – VisionVibe
- 1126 – PixelPal
- 1127 – FreezeFrame

- 1128 – AngleAttack
- 1129 – SnapStash
- 1130 – FrameFlair
- 1131 – SmileSnaps
- 1132 – ImageImpact
- 1133 – ZoomZest
- 1134 – ClickCraze
- 1135 – PixelPassion
- 1136 – ShootSmart
- 1137 – CaptionChamp.
- 1138 • grocery_and_gourmet_food.:
- 1139 – Nutella
- 1140 – Doritos
- 1141 – Hellmann’s Mayonnaise
- 1142 – Campbell’s Soup
- 1143 – Ritz Crackers
- 1144 – Quaker Oats
- 1145 – Ben & Jerry’s Ice Cream
- 1146 – Tostitos Salsa
- 1147 – Goldfish Crackers
- 1148 – Red Bull Energy Drink
- 1149 – McCormick Spices
- 1150 – Crystal Light Drink Mix
- 1151 – Funyuns Onion Rings
- 1152 – Skippy Peanut Butter
- 1153 – Heinz Ketchup
- 1154 – Tabasco Hot Sauce
- 1155 – Hershey’s Chocolate Syrup
- 1156 – Nescafe Coffee
- 1157 – Kraft Macaroni & Cheese
- 1158 – Gatorade Sports Drink
- 1159 • baby.:
- 1160 – Baby Swing
- 1161 – Diaper Genie
- 1162 – Milk Warmer
- 1163 – Baby Carrier
- 1164 – Car Seat
- 1165 – Baby Monitor
- 1166 – Baby Food Maker
- 1167 – Nursery Glider
- 1168 – Changing Table
- 1169 – Baby Bouncer
- 1170 – Playpen
- 1171 – Teething Rings
- 1172 – Baby Wipes Dispenser
- 1173 – Baby Bath Tub

- 1174 – Bibs
- 1175 – Baby Blankets
- 1176 – Pacifier Clip
- 1177 – Baby Sling
- 1178 – Baby Napper
- 1179 – Moses Basket
- 1180 • outdoor_living.:
 - 1181 – Sunbrella
 - 1182 – Weber
 - 1183 – Keter
 - 1184 – Trex
 - 1185 – Solaira
 - 1186 – Tropitone
 - 1187 – Bimini Solutions
 - 1188 – La-Z-Boy Outdoor
 - 1189 – Suncast
 - 1190 – Beltwood
 - 1191 – Quikrete
 - 1192 – Cosco Outdoor Living
 - 1193 – Anova Furnishings
 - 1194 – Duramax
 - 1195 – US Polymers
 - 1196 – Ostrich Chairs
 - 1197 – Carefree of Colorado
 - 1198 – Tuff Coat
 - 1199 – Fire Sense
 - 1200 – Heritage Patios.

1201 **I.3 Reddit Attributes**

1202 **I.3.1 Resources**

1203 We randomly select 5 categories in Reddit dataset and display the corresponding resource attributes
 1204 for each category:

- 1205 • buddy_crossing.:
 - 1206 – Meetup.com - a website that allows you to find and join groups of individuals with
 1207 similar interests in your local area, including hiking, book clubs, and social events.
 - 1208 – The Buddy System: Understanding Mental Illness and Addiction - a book that explores
 1209 the biology of addiction and provides a guide for friends and family members of
 1210 individuals struggling with these issues.
 - 1211 – Lynda.com - a subscription-based online learning platform that provides courses on a
 1212 variety of subjects including computer programming, business, web design, and more.
 - 1213 – Codecademy.com - an interactive online platform that teaches coding skills for free or
 1214 through a subscription.
 - 1215 – Khan Academy - a nonprofit organization that provides free online courses in a wide
 1216 range of subjects including math, science, and humanities to learners of all ages.
 - 1217 – Duolingo - a language-learning app that is available for free on the App Store and
 1218 Google Play, offering courses in a variety of languages including Spanish, French, and
 1219 German.

- 1220 – MindBody App - a mobile app that helps users find and book local fitness, wellness,
1221 and beauty activities such as yoga classes, massages, and haircuts.
- 1222 – Headspace - a meditation app that offers guided meditation courses to help users reduce
1223 stress and improve focus.
- 1224 – The Knot - a website that provides tools and resources for wedding planning, including
1225 a Wedding Website Builder, guest list tracker, and registry management.
- 1226 – Khan Academy - a nonprofit organization that provides free online courses in a wide
1227 range of subjects including math, science, and humanities to learners of all ages.
- 1228 – Others resource for buddy_crossing.
- 1229 • the_division.:
 - 1230 – Division Builds - A subreddit dedicated to sharing and discussing various builds used
1231 in The Division.
 - 1232 – Division Zone - A website with extensive information on game mechanics, gear, and
1233 other important gameplay aspects.
 - 1234 – The Division Discord - A community-run Discord server where players can connect
1235 and find groups to play with.
 - 1236 – The Division Wiki - A comprehensive wiki with guides, tips, and information on
1237 everything related to The Division.
 - 1238 – Skill-Up's YouTube channel - A popular YouTuber who provides detailed analysis and
1239 reviews of The Division's updates and patches.
 - 1240 – MarcoStyle's YouTube channel - Another popular YouTuber who provides in-depth
1241 analysis and guides for The Division's gameplay and mechanics.
 - 1242 – The Division LFG - A website where players can find groups to play with, organize
1243 events, and share their experiences.
 - 1244 – The Division Zone Map - An interactive map that allows players to find important
1245 points of interest, loot, and other useful resources.
 - 1246 – The Division 2 subreddit - A community-run subreddit for the sequel, The Division 2,
1247 where players can share their experiences and discuss the game.
 - 1248 – Others resource for the_division.
- 1249 • roblox.:
 - 1250 – Roblox Wiki (https://roblox.fandom.com/wiki/Main_Page)
 - 1251 – Roblox Developer Forum (<https://devforum.roblox.com/>)
 - 1252 – Ultimate Guide to Making Your First Game on Roblox
1253 (<https://medium.com/@Piranhari/ultimate-guide-to-making-your-first-game-on-roblox-part-1-f1fc63abf7>)
 - 1254 – Roblox Blog (<https://blog.roblox.com/>)
 - 1255 – Roblox Studio Tutorials (<https://www.youtube.com/playlist?list=PLuEQ5BB-Z1SgeZTAAq2w1K3kUfQ-yLEOj>)
 - 1256 – The Roblox Developer Hub (<https://developer.roblox.com/en-us/>)
 - 1257 – Top 10 Roblox Games (<https://www.techjunkie.com/top-10-best-roblox-games/>)
 - 1258 – Roblox Discord Server (<https://discord.gg/roblox>)
 - 1259 – Roblox Support (<https://en.help.roblox.com/hc/en-us>)
 - 1260 – Top Roblox Youtubers to Follow (<https://www.gamertweak.com/top-roblox-youtubers-to-follow/>)
 - 1261 – Others resource for roblox.
- 1262 • whats_that_book.:
 - 1263 – Goodreads - A social platform for book lovers where users can search for books, create
1264 bookshelves, and write reviews.

- 1268 – LibraryThing - A community-driven cataloging website where users can create and
- 1269 share their personal book collections.
- 1270 – AbeBooks - An online marketplace for rare and out-of-print books, as well as other
- 1271 antique or collectible items.
- 1272 – Shelfari - An online book club where users can share book recommendations and read
- 1273 reviews from others.
- 1274 – Project Gutenberg - A digital library of freely available public domain books.
- 1275 – Paperback Swap - A book trading community where users can exchange books with
- 1276 others across the US.
- 1277 – Goodreads Librarians Group - A community of Goodreads users who help with book
- 1278 cataloging, including identifying books from incomplete information.
- 1279 – Book Riot - A website featuring book reviews and book-related news, with an emphasis
- 1280 on diverse and underrepresented voices.
- 1281 – The New York Times Book Review - A renowned weekly publication featuring book
- 1282 reviews, author interviews, and literary criticism.
- 1283 – Others resource for whats_that_book.
- 1284 • pokemongo_friends.:
 - 1285 – Pokemon GO Hub: A comprehensive website dedicated to news, guides, and analysis
 - 1286 on Pokemon GO.
 - 1287 – The Silph Road Subreddit: A community-run subreddit dedicated to research and
 - 1288 analysis of Pokemon GO mechanics.
 - 1289 – Poke Assistant: A website that offers a range of tools to help you optimize your
 - 1290 Pokemon GO experience, including IV calculators and gym battle simulations.
 - 1291 – The Trainer Club: A YouTube channel that provides daily updates, news, and tips for
 - 1292 Pokemon GO trainers.
 - 1293 – Gotta Catch 'Em All: A Facebook group where you can connect with other Pokemon
 - 1294 GO players and coordinate raid battles and other activities.
 - 1295 – Reddit's r/PokemonGOFriends Subreddit: A community of players looking for friends
 - 1296 to exchange gifts and share invites for raids.
 - 1297 – The PokeMap: A website that allows you to find nearby Pokemon on a map in real-time.
 - 1298 – Poke Genie: An app that automatically calculates IVs and other stats for your Pokemon,
 - 1299 saving you time and headaches.
 - 1300 – Pokemon GO Gamepress: A website that offers detailed breakdowns and analysis of
 - 1301 Pokemon, movesets, and other game mechanics.
 - 1302 – The Go Ranger App: An app that helps you plan your raids and battles, with intuitive
 - 1303 mapper tools and filters to help you find the Pokemon you're looking for.
 - 1304 – Others resource for pokemongo_friends.

1305 I.3.2 Experience

1306 We randomly select 5 categories in Reddit dataset and display the corresponding experience attributes
 1307 for each category:

- 1308 • build_a_pc.:
 - 1309 – DIY PC Builds: Sharing personal experiences and success stories of building custom
 - 1310 PCs, discussing component choices, troubleshooting, and performance optimizations.
 - 1311 – Budget-Friendly Builds: Discussing experiences with building PCs on a tight budget,
 - 1312 sharing cost-saving tips, and recommendations for budget-friendly components.
 - 1313 – Cable Management: Sharing personal experiences and tips for effective cable manage-
 - 1314 ment in PC builds, discussing cable routing techniques and showcasing clean build
 - 1315 aesthetics.

- 1316 – RGB Lighting: Discussing experiences with RGB lighting setups in PC builds, sharing
- 1317 recommendations for RGB components, software customization, and lighting effects.
- 1318 – Troubleshooting Builds: Sharing experiences and tips for troubleshooting common
- 1319 issues in PC builds, helping fellow builders diagnose and solve hardware or software
- 1320 problems.
- 1321 – Silent and Quiet PC Builds: Discussing experiences and recommendations for building
- 1322 silent or quiet PCs, focusing on noise reduction techniques and quiet component
- 1323 choices.
- 1324 – Workstation Builds: Sharing experiences and insights into building PCs for professional
- 1325 workloads, such as video editing, 3D rendering, programming, and graphic design.
- 1326 – Water-Cooling Adventures: Sharing experiences and insights into custom water-cooling
- 1327 loops, discussing the challenges, benefits, and performance improvements achieved.
- 1328 – Unique and Custom Builds: Showcasing and discussing unique and custom PC builds,
- 1329 including themed builds, custom cases, or exotic cooling solutions.
- 1330 – Build Planning and Component Selection: Discussing experiences with planning PC
- 1331 builds, researching and selecting components, considering compatibility, and balancing
- 1332 performance and budget.
- 1333 – Modding and Case Customization: Sharing experiences with PC case modding and
- 1334 customization, discussing techniques, materials, and showcasing personal projects.
- 1335 – Compact and Small Form Factor Builds: Discussing experiences with building compact
- 1336 or small form factor PCs, sharing recommendations for mini-ITX cases, cooling
- 1337 solutions, and component choices.
- 1338 – Home Server and NAS Builds: Sharing experiences and insights into building home
- 1339 servers and network-attached storage (NAS) systems, discussing storage options, soft-
- 1340 ware, and data management.
- 1341 – Multimonitor Setups: Discussing experiences with multimonitor setups, sharing tips
- 1342 for optimizing productivity and gaming experiences across multiple displays.
- 1343 – PC Gaming Peripherals: Sharing experiences and recommendations for gaming pe-
- 1344 ripherals, such as keyboards, mice, monitors, and headsets, discussing features and
- 1345 personal preferences.
- 1346 • `summon_sign`.:
 - 1347 – Sunbro Covenant: Embracing the Sunbro covenant and assisting fellow players with
 - 1348 jolly cooperation, earning sunlight medals and praising the sun together.
 - 1349 – Fashion Souls: Sharing and showcasing unique and fashionable character builds, armor
 - 1350 sets, and weapon combinations for aesthetic enjoyment.
 - 1351 – Covenant Experiences: Sharing experiences and strategies related to various in-game
 - 1352 covenants, such as the Darkwraiths, Blades of the Darkmoon, or Forest Hunters.
 - 1353 – Community Creations: Showcasing community-created content, such as fan art, videos,
 - 1354 or fan fiction, celebrating the creativity and talent within the Summon Sign community.
 - 1355 – Lore-friendly Builds: Discussing and sharing character builds that are aligned with
 - 1356 specific characters or factions within the game’s lore, adding immersion and roleplaying
 - 1357 elements.
 - 1358 – Community Appreciation: Expressing gratitude and appreciation for the community,
 - 1359 developers, and the overall enjoyment derived from the Dark Souls series and the
 - 1360 cooperative multiplayer experiences.
- 1361 • `xbox`.:
 - 1362 – Xbox One exclusive games such as Halo 5, Forza Horizon 4, and Gears of War 4
 - 1363 – Xbox One media and entertainment apps such as Netflix and Hulu
 - 1364 – memorable gaming moments or achievements on the Xbox console.
 - 1365 – Purchase Xbox One online.

- 1366 – Xbox Kinect motion sensor accessory
- 1367 – Xbox Play Anywhere program
- 1368 – Other Experience of Xbox
- 1369 • pittsburgh.:
 - 1370 – Visit the Andy Warhol Museum
 - 1371 – Watch a Steelers football game at Heinz Field
 - 1372 – Explore the Carnegie Museum of Natural History
 - 1373 – Ride to the top of Mount Washington on the Duquesne Incline
 - 1374 – Take a leisurely stroll through Phipps Conservatory and Botanical Gardens
 - 1375 – Experience the history of the city at the Senator John Heinz History Center
 - 1376 – Tour the University of Pittsburgh campus
 - 1377 – Attend a performance at the Benedum Center for the Performing Arts
 - 1378 – Take a walk along the Three Rivers Heritage Trail
 - 1379 – Taste pierogies, kielbasa, and other traditional Pittsburgh foods
 - 1380 – Admire the architecture of the Cathedral of Learning
 - 1381 – Explore the Pittsburgh Zoo and PPG Aquarium
 - 1382 – Gaze in awe at the exhibits in the Carnegie Science Center
 - 1383 – Visit the National Aviary and get up close with tropical birds
 - 1384 – Check out the local art scene on Penn Avenue in the Garfield neighborhood
 - 1385 – Attend the Three Rivers Regatta, Pittsburgh’s largest annual summer event
 - 1386 – Take a bike ride on the Great Allegheny Passage trail
 - 1387 – Ride the roller coasters at Kennywood Amusement Park
 - 1388 – Discover the nightlife in the South Side neighborhood
 - 1389 – Go shopping at the Strip District markets for locally-made goods and fresh produce.
 - 1390 – Others experience for pittsburgh.
- 1391 • metal_gear_solid.:
 - 1392 – Tactical Weapon Customization: Experimenting with various weapons, attachments, and equipment to tailor loadouts to different mission objectives and playstyles.
 - 1393 – Character Development: Witnessing the growth and development of iconic characters such as Solid Snake, Big Boss, or Raiden throughout their respective story arcs.
 - 1394 – Stealthy Takedowns: Executing silent and non-lethal takedowns, utilizing tranquilizer darts, chokeholds, or sneaking up on enemies from behind.
 - 1395 – Tactical Planning: Strategizing mission approaches, analyzing enemy patrols, setting traps, and utilizing distractions to gain the upper hand.
 - 1396 – Memorable Characters: Developing connections with unique and memorable characters, such as Otacon, Meryl, Revolver Ocelot, or The Boss.
 - 1397 – Stealthy Gadgets: Utilizing gadgets and tools, such as the cardboard box, thermal goggles, or the iconic Solid Eye, to gain advantages during missions.
 - 1398 – Emotional Storytelling: Experiencing emotionally impactful moments within the narrative, exploring themes of loss, betrayal, loyalty, and the human cost of warfare.
 - 1399 – Espionage Tactics: Participating in undercover missions, gathering intelligence, infiltrating enemy bases, and sabotaging their operations.
 - 1400 – 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.
 - 1401 – 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.
 - 1402 – Others experience for metal_gear_solid.

1415 I.4 StackExchange Attributes

1416 I.4.1 Scenarios

1417 We randomly select 5 categories in StackExchange dataset and display the corresponding scenario
1418 attributes for each category:

- 1419 • multiplayer.:
 - 1420 – Cheating/hacking in online games
 - 1421 – Inappropriate player behavior
 - 1422 – Unbalanced game mechanics
 - 1423 – Difficulty connecting to multiplayer servers
 - 1424 – Matchmaking errors
 - 1425 – Unresponsive or laggy gameplay
 - 1426 – Glitches in gameplay affecting online matches
 - 1427 – Difficulty finding players to match with
 - 1428 – Balancing player skills in matchmaking
 - 1429 – Disconnects and dropped connections mid-game
 - 1430 – Cross-platform compatibility issues
 - 1431 – In-game communication problems
 - 1432 – Difficulty managing and moderating game servers
 - 1433 – Addressing griefing and trolling in multiplayer games
 - 1434 – Managing player accounts and login systems
 - 1435 – Implementing or improving anti-cheat measures
 - 1436 – Community feedback and feature requests
 - 1437 – Addressing game-breaking exploits
 - 1438 – Ensuring fair and accurate reporting of player statistics
 - 1439 – Addressing server crashes and downtime.
- 1440 • terrain.:
 - 1441 – Error in generative algorithms for creating terrain
 - 1442 – Difficulty in implementing procedural terrain generation in a specific game engine
 - 1443 – Inconsistencies in terrain generation across different devices
 - 1444 – Issues with realism in terrain generation algorithms
 - 1445 – Difficulty in implementing terrain physics and collision detection
 - 1446 – Terrain rendering issues on low-spec hardware
 - 1447 – Incompatibility between terrain generation and map or level editors
 - 1448 – Optimization of terrain generation algorithms for speed and memory usage
 - 1449 – Unwanted artifacts and glitches in terrain mesh generation
 - 1450 – Compatibility issues between terrain generation algorithms and game engine systems
 - 1451 – Difficulty in creating realistic terrain textures and materials
 - 1452 – Inaccuracy of terrain elevation generation in certain geographic regions
 - 1453 – Difficulty in implementing terrain deformation mechanics
 - 1454 – Poor performance with large-scale terrain rendering and generation
 - 1455 – Unwanted noise and roughness in generated terrain meshes
 - 1456 – Compatibility issues between terrain generation and asset importation pipelines
 - 1457 – Inaccuracy of terrain heightmap data due to low-quality input data sources
 - 1458 – Difficulty in handling multi-layer terrain materials and textures
 - 1459 – Poor performance with dynamic terrain generation and updates

- 1460 – Issues with biome and climate-based terrain generation.
- 1461 • rendering.:
 - 1462 – Difficulty creating realistic hair and fur in rendering software.
 - 1463 – Debugging issues with transparent materials in a 3D rendering engine.
 - 1464 – Crashes or slow performance when rendering large scenes in real time.
 - 1465 – Trouble with anti-aliasing and other graphics optimization techniques.
 - 1466 – Struggle with optimizing rendering quality on lower-end hardware.
 - 1467 – Difficulty incorporating custom shaders into a game engine or rendering pipeline.
 - 1468 – Figuring out how to use the latest rendering features in a legacy project.
 - 1469 – Issues with rendering dynamic lighting in real time, without pre-baking.
 - 1470 – Optimizing flicker or aliasing issues caused by fast-moving objects in a scene.
 - 1471 – Solving glitches or crashes caused by malfunctioning GPU drivers.
 - 1472 – Difficulty in rendering complex ocean or water simulations.
 - 1473 – Troubleshooting issues with volumetric rendering in a 3D engine.
 - 1474 – Finding the optimal rendering settings for a particular 3D model or scene.
 - 1475 – Figuring out optimal texturing and lighting in a photorealistic rendering.
 - 1476 – Creating procedural textures and materials in a game engine.
 - 1477 – Debugging flicker issues caused by overlapping or intersecting geometry.
 - 1478 – Difficulty in rendering realistic motion blur in 3D animation.
 - 1479 – Solving imbalanced lighting in a photorealistic rendering of a room or scene.
 - 1480 – Finding the optimal rendering settings for VR or AR applications.
 - 1481 – Debugging issues with inaccurate or glitchy global illumination in a scene.
- 1482 • procedural-generation.:
 - 1483 – Improving the efficiency of procedural generation algorithms in Python.
 - 1484 – Troubleshooting issues with memory usage in large-scale procedural generation projects.
 - 1485 – Debugging issues with randomized content generation in procedural levels.
 - 1487 – Implementing procedural generation techniques in C++ for game development.
 - 1488 – Exploring the potential of machine learning in procedural content generation.
 - 1489 – Optimizing the generation of complex 3D models using procedural algorithms.
 - 1490 – Managing complexity and maintaining consistency in procedurally generated game worlds.
 - 1491 – Addressing issues with procedural generation of text-based content, such as dialogue or item descriptions.
 - 1492 – Developing tools to aid in the creation and testing of procedural generation algorithms.
 - 1493 – Balancing the need for randomness with player expectations for fairness and balance.
 - 1494 – Addressing issues with the procedural generation of music and sound effects.
 - 1495 – Improving the visual quality of procedurally generated game assets.
 - 1496 – Exploring ethical concerns around the use of AI in procedural content generation.
 - 1497 – Developing procedural generation techniques for non-linear narratives or branching storylines.
 - 1498 – Improving the procedural generation of inhabited environments, such as procedurally generated NPCs.
 - 1499 – Addressing issues with the procedural generation of terrain features such as rivers and mountains.
 - 1500 – Implementing procedural generation techniques for user-generated content.
 - 1501 – Supporting multithreaded execution in procedural generation algorithms.

- 1507 – Ensuring procedural generation techniques are compatible with various game engines
- 1508 and frameworks.
- 1509 – Improving the scalability of procedural generation algorithms for use in multiplayer
- 1510 games.
- 1511 • networking.:
- 1512 – Difficulty in troubleshooting network connection issues on a Linux system
- 1513 – Configuring a wireless access point for a large office space
- 1514 – Implementing load balancing across multiple servers in a network
- 1515 – Optimizing network throughput to reduce latency in a gaming environment
- 1516 – Implementing firewall rules to block unauthorized access to a network
- 1517 – Troubleshooting DNS resolution issues on a Windows server
- 1518 – Designing and implementing a secure VPN connection
- 1519 – Setting up a network file server for shared access among multiple clients
- 1520 – Configuring SNMP to monitor network traffic and utilization
- 1521 – Designing a network topology for a large enterprise with multiple locations
- 1522 – Troubleshooting issues with Ethernet switches in a data center environment
- 1523 – Implementing QoS to prioritize network traffic for critical applications
- 1524 – Configuring NAT and PAT to enable internet access for multiple devices
- 1525 – Setting up and configuring VLANs to segment a network
- 1526 – Troubleshooting issues with network printers in an office environment
- 1527 – Configuring routing protocols in a large network
- 1528 – Securing wireless access points to prevent unauthorized access
- 1529 – Troubleshooting issues with VPN connection stability and speed
- 1530 – Implementing network virtualization with virtual LANs and virtual switches
- 1531 – Designing and implementing an effective network security strategy to prevent data
- 1532 breaches.

1533 I.5 ArXiv Attributes

1534 I.5.1 Subtopics

1535 We randomly select 5 categories in arXiv dataset and display the corresponding product brand
 1536 attributes for each category:

- 1537 • hardware_architecture.:
- 1538 – Processor design
- 1539 – Memory architecture
- 1540 – Input/output (I/O) systems
- 1541 – Computer networks
- 1542 – Digital signal processing
- 1543 – Microcontrollers
- 1544 – Embedded systems
- 1545 – Real-time systems
- 1546 – Graphics processing units (GPUs)
- 1547 – High-performance computing (HPC)
- 1548 – Computer architecture for mobile devices
- 1549 – Wearable technology hardware design
- 1550 – Secure hardware systems
- 1551 – Fault-tolerant hardware design

- 1552 – Quantum computing architecture
- 1553 – Bio-inspired computing systems
- 1554 – Nanoscale computing systems
- 1555 – Reconfigurable hardware systems
- 1556 – Internet of Things (IoT) devices
- 1557 – Green computing and sustainability in hardware design
- 1558 • genomics.:
 - 1559 – Genome editing and CRISPR technology
 - 1560 – Epigenomics and epigenetic modifications
 - 1561 – Pharmacogenomics and personalized medicine
 - 1562 – Functional genomics and transcriptomics
 - 1563 – Metagenomics and microbiome analysis
 - 1564 – Genomic epidemiology and infectious disease
 - 1565 – Cancer genomics and tumor heterogeneity
 - 1566 – Structural genomics and protein structure prediction
 - 1567 – Plant genomics and crop breeding
 - 1568 – Animal genomics and livestock improvement
 - 1569 – Comparative genomics and evolution
 - 1570 – Genomic diversity and population genetics
 - 1571 – Next-generation sequencing and high-throughput analysis
 - 1572 – Single-cell genomics and cellular heterogeneity
 - 1573 – Ethical, legal, and social implications of genomics
 - 1574 – Cryptography and secure data sharing in genomics
 - 1575 – Big data analytics and data mining in genomics
 - 1576 – Immune system genomics and immunotherapy
 - 1577 – Genomics and aging research
 - 1578 – Gene therapy and gene regulation.
- 1579 • machine_learning.:
 - 1580 – Text generation
 - 1581 – Natural language understanding for chatbots
 - 1582 – Sentiment analysis and opinion mining
 - 1583 – Text summarization and keyword extraction
 - 1584 – Machine translation
 - 1585 – Named entity recognition and entity linking
 - 1586 – Dialogue systems and conversational agents
 - 1587 – Cross-lingual and Multilingual NLP
 - 1588 – Text-to-speech systems
 - 1589 – Phonetics and phonology in computational linguistics
 - 1590 – Grammatical error detection and correction
 - 1591 – Speech recognition and acoustic modeling
 - 1592 – Semantic role labeling
 - 1593 – Discourse analysis and coherence modeling
 - 1594 – Lexical semantics and word sense disambiguation
 - 1595 – Computational lexicography and machine-readable dictionaries
 - 1596 – Language Modeling
 - 1597 – question answering

- 1598 – Language resources and corpora
- 1599 – Computational sociolinguistics and dialectology.
- 1600 • number_theory.:
 - 1601 – Prime numbers
 - 1602 – Diophantine equations
 - 1603 – Modular arithmetic
 - 1604 – Cryptography
 - 1605 – Continued Fractions
 - 1606 – Pell’s Equation
 - 1607 – Fermat’s Last Theorem
 - 1608 – Algebraic Number Theory
 - 1609 – Riemann Hypothesis
 - 1610 – Arithmetic Geometry
 - 1611 – Quadratic Forms
 - 1612 – L-Functions
 - 1613 – Automorphic Forms
 - 1614 – Galois Theory
 - 1615 – Ramsey Theory
 - 1616 – Distribution of Prime Numbers
 - 1617 – Number Theory in Cryptography
 - 1618 – Summation Formulas
 - 1619 – Gaussian Integers
 - 1620 – The Goldbach Conjecture
- 1621 • geophysics.:
 - 1622 – Seismic imaging
 - 1623 – Earthquake prediction
 - 1624 – Geothermal energy
 - 1625 – Volcanic eruptions
 - 1626 – Plate tectonics
 - 1627 – Geomagnetism
 - 1628 – Paleomagnetism
 - 1629 – Geophysical surveying
 - 1630 – Geophysical fluid dynamics
 - 1631 – Gravity measurements
 - 1632 – Rock physics
 - 1633 – Crustal deformation
 - 1634 – Geomorphology
 - 1635 – Mineral exploration
 - 1636 – Earth structure modeling
 - 1637 – Geodetic techniques
 - 1638 – Hydrogeophysics
 - 1639 – Earth modeling
 - 1640 – Electrical geophysics
 - 1641 – Remote sensing geophysics

1642 **I.5.2 Techniques**

1643 We randomly select 5 categories in the arXiv dataset and display the corresponding attributes for each
1644 category:

- 1645 • hardware_architecture.:
 - 1646 – Binary Decision Diagrams (BDDs) for circuit optimization
 - 1647 – Probabilistic CMOS (PCMOs) for energy-efficient computing
 - 1648 – Boundary Scan Architecture (BSA) for testing and fault diagnosis
 - 1649 – Clock gating for power reduction in synchronous circuits
 - 1650 – Memristive networks for neuromorphic computing
 - 1651 – Approximate computing for energy-efficient digital processing
 - 1652 – Verilog Hardware Description Language (HDL) for circuit design and simulation
 - 1653 – Reversible logic circuits for low power consumption
 - 1654 – High-level synthesis (HLS) for synthesizing hardware from software descriptions
 - 1655 – Phase Change Memory (PCM) for high-density non-volatile storage.
- 1656 • genomics.:
 - 1657 – Genome assembly and annotation using hybrid approaches.
 - 1658 – Comparative genomics for analyzing evolutionary relationships between genomes.
 - 1659 – Differential gene expression analysis using RNA sequencing data.
 - 1660 – Metagenomics for studying the microbial communities in different environments.
 - 1661 – Epigenetic analysis for understanding gene regulation.
 - 1662 – Network analysis for identifying gene interactions and pathways.
 - 1663 – Structural variation analysis for detecting genomic rearrangements.
 - 1664 – Functional genomics for studying gene function and pathway regulation.
 - 1665 – Genome-wide association studies for identifying genetic variants associated with
 - 1666 complex traits.
 - 1667 – High-throughput screening methods for identifying genes involved in specific biological
 - 1668 processes.
- 1669 • computation_and_language.:
 - 1670 – Word Embeddings
 - 1671 – Recurrent Neural Networks (RNNs)
 - 1672 – Attention Mechanism
 - 1673 – Transformer Model
 - 1674 – Dependency Parsing
 - 1675 – Topic Modeling:
 - 1676 – Machine Translation
 - 1677 – Corpus Analysis
 - 1678 – Discourse Analysis
 - 1679 – Pragmatics
- 1680 • number_theory:
 - 1681 – Primality testing using elliptic curves
 - 1682 – Continued fraction factorization method
 - 1683 – Algorithm for solving Diophantine equations
 - 1684 – Quadratic sieve algorithm for integer factorization
 - 1685 – Pollard rho algorithm for integer factorization
 - 1686 – Digital sum subtraction method for computing discrete logarithm

- 1687 – Fermat’s method for factorization of primes
- 1688 – Chinese remainder algorithm for solving modular equations
- 1689 – Exponential-sum algorithm for computing in algebraic number fields
- 1690 – Generalized Ramanujan-Selberg formula for counting integer points on algebraic
- 1691 varieties.
- 1692 • geophysics.:
 - 1693 – Seismic attribute interpretation
 - 1694 – Full waveform inversion
 - 1695 – Gravity inversion
 - 1696 – Spherical geometries
 - 1697 – Ground penetrating radar imaging
 - 1698 – Time-lapse reservoir monitoring
 - 1699 – Electrical resistivity tomography
 - 1700 – Joint inversion of geophysical data
 - 1701 – Radiometric dating
 - 1702 – Geomagnetic field modeling

1703 **I.6 AG News Attributes**

1704 **I.6.1 Subtopics**

1705 We randomly select 5 categories in AG News dataset and display the corresponding subtopic attributes
 1706 for each category:

- 1707 • business:
 - 1708 – Corporate earnings and financial reports
 - 1709 – Stock market updates and analysis
 - 1710 – Mergers and acquisitions
 - 1711 – Business regulations and policies
 - 1712 – Startups and entrepreneurship
 - 1713 – Industry trends and forecasts
 - 1714 – Economic indicators and market trends
 - 1715 – Business strategies and management practices
 - 1716 – Corporate governance and ethics
 - 1717 – Consumer behavior and market research
 - 1718 – Business leadership and executive profiles
 - 1719 – Banking and finance industry updates
 - 1720 – Energy and sustainability in business
 - 1721 – Retail and e-commerce trends
 - 1722 – Real estate and property market updates
 - 1723 – Business disruptions and crisis management
 - 1724 – Corporate social responsibility and sustainability initiatives
- 1725 • sci_tech:
 - 1726 – Artificial intelligence
 - 1727 – Robotics
 - 1728 – Quantum computing
 - 1729 – Biotechnology
 - 1730 – Nanotechnology

- 1731 – Internet of Things
- 1732 – Renewable energy
- 1733 – Virtual reality
- 1734 – Augmented reality
- 1735 – Cybersecurity
- 1736 – Genetic engineering
- 1737 – Big data
- 1738 – Autonomous vehicles
- 1739 – 3D printing
- 1740 – Blockchain technology
- 1741 – Bioinformatics
- 1742 – Machine learning
- 1743 – Biomedical engineering
- 1744 – Clean technology
- 1745 • sports:
 - 1746 – Soccer
 - 1747 – Basketball
 - 1748 – Baseball
 - 1749 – Tennis
 - 1750 – Golf
 - 1751 – Cricket
 - 1752 – Rugby
 - 1753 – Athletics
 - 1754 – Formula 1
 - 1755 – Olympics
 - 1756 – Boxing
 - 1757 – Swimming
 - 1758 – Volleyball
 - 1759 – Ice hockey
 - 1760 – American football
 - 1761 – Cycling
 - 1762 – Motorsports
 - 1763 – Martial arts
 - 1764 – Horse racing
 - 1765 – Surfing
- 1766 • world:
 - 1767 – International politics and diplomacy
 - 1768 – Global conflicts and war
 - 1769 – Terrorism and security threats
 - 1770 – Human rights issues and social justice movements
 - 1771 – Migration and refugee crises
 - 1772 – Climate change and environmental policies
 - 1773 – Global health crises and pandemics
 - 1774 – Natural disasters and emergencies
 - 1775 – Cross-border crime and corruption
 - 1776 – Cultural and social developments worldwide

- 1777 – Geopolitical tensions and territorial disputes
- 1778 – International aid and development efforts
- 1779 – Humanitarian crises and relief efforts
- 1780 – Cultural heritage preservation and promotion
- 1781 – International collaborations and partnerships

1782 **I.7 SST-2 Attributes**

1783 **I.7.1 Subtopics**

1784 We randomly select 5 categories in SST-2/IMDB (movie review) dataset and display the corresponding
 1785 subtopic attributes for each category:

- 1786 • positive:
 - 1787 – Compelling Storyline: A strong and engaging narrative that captures the audience’s
 - 1788 attention from beginning to end.
 - 1789 – Well-Developed Characters: Memorable and relatable characters that evoke emotions
 - 1790 and drive the story forward.
 - 1791 – Skillful Direction: Effective direction that showcases the filmmaker’s vision, ensuring
 - 1792 cohesive storytelling and engaging visual elements.
 - 1793 – Excellent Acting: Convincing performances from the cast that bring the characters to
 - 1794 life and immerse the audience in the story.
 - 1795 – Cinematography: Expertly captured visuals, including the use of framing, lighting,
 - 1796 and camera movements, to enhance the storytelling and create a visually appealing
 - 1797 experience.
 - 1798 – Engaging Dialogue: Well-written dialogue that is natural, meaningful, and contributes
 - 1799 to character development and plot progression.
 - 1800 – Sound Design and Music: Thoughtful and immersive sound design, including sound
 - 1801 effects and a well-curated soundtrack or original score, that enhances the overall
 - 1802 cinematic experience.
 - 1803 – Production Design: Attention to detail in creating visually appealing and authentic
 - 1804 sets, costumes, and overall aesthetics that contribute to the film’s atmosphere and
 - 1805 world-building.
 - 1806 – Editing: Skillful editing that maintains a good pace, effectively transitions between
 - 1807 scenes, and enhances the overall flow and impact of the story.
 - 1808 – Emotional Impact: A movie that evokes emotions, whether it be through humor, drama,
 - 1809 suspense, or other means, leaving a lasting impression on the audience.
- 1810 • negative:
 - 1811 – Weak Plot: A poorly developed or uninteresting storyline that fails to engage the
 - 1812 audience.
 - 1813 – Lackluster Performances: Unconvincing or uninspired performances by the actors that
 - 1814 fail to bring the characters to life.
 - 1815 – Poor Production Quality: Subpar production values, including low-quality visuals,
 - 1816 amateurish cinematography, and weak special effects.
 - 1817 – Incoherent Storytelling: Confusing or disjointed narrative structure that makes it
 - 1818 difficult to follow or understand the plot.
 - 1819 – Unmemorable Characters: Underdeveloped or forgettable characters that fail to resonate
 - 1820 with the audience.
 - 1821 – Weak Soundtrack: A forgettable or poorly composed soundtrack that fails to enhance
 - 1822 the mood or add depth to the movie.
 - 1823 – Poor Dialogue: Uninteresting or poorly written dialogues that fail to engage or resonate
 - 1824 with the audience.

- 1825 – Disjointed Atmosphere: A lack of coherence or consistency in creating an immersive
- 1826 and believable world for the viewers.
- 1827 – Unresolved Plotlines: Loose ends or unresolved plotlines that leave the audience feeling
- 1828 unsatisfied or confused.
- 1829 – Lack of Entertainment Value: A movie that fails to deliver an enjoyable or engaging
- 1830 experience for the audience, leaving them feeling bored or uninterested.

1831 **I.7.2 Descriptive Details**

1832 We use movie genres as the characteristics of movies, and the attributes are listed as follows:

- 1833 • Action
- 1834 • Drama
- 1835 • Comedy
- 1836 • Thriller
- 1837 • Romance
- 1838 • Horror
- 1839 • Adventure
- 1840 • Science Fiction
- 1841 • Fantasy
- 1842 • Animation

1843 **I.8 Yelp Attributes**

1844 **I.8.1 Subtopics**

1845 We randomly select 5 categories in Yelp review dataset and display the corresponding subtopic

1846 attributes for each category:

- 1847 • positive:
 - 1848 – Quality of Food: The taste, flavor, and presentation of the dishes.
 - 1849 – Fresh Ingredients: The use of fresh and high-quality ingredients in the preparation of
 - 1850 the food.
 - 1851 – Menu Variety: A diverse range of options catering to different dietary preferences and
 - 1852 restrictions.
 - 1853 – Presentation: The visually appealing presentation of the dishes.
 - 1854 – Service: Attentive, friendly, and prompt service from the restaurant staff.
 - 1855 – Value for Money: Offering good quality and portion sizes at reasonable prices.
 - 1856 – Cleanliness: A clean and well-maintained dining area, including tables, utensils, and
 - 1857 restrooms.
 - 1858 – Special Dietary Accommodations: Catering to specific dietary needs such as vegetarian,
 - 1859 vegan, gluten-free, etc.
 - 1860 – Unique and Creative Dishes: Offering innovative and creative dishes that stand out.
 - 1861 – Efficient Operations: Smooth and well-coordinated operations to minimize waiting
 - 1862 times and delays.
- 1863 • negative:
 - 1864 – Poor Service: Slow or inattentive service from the restaurant staff. Unfriendly Staff:
 - 1865 Rude or unhelpful behavior from the restaurant staff.
 - 1866 – Long Waiting Times: Excessive waiting times for a table or food.
 - 1867 – Incorrect Orders: Receiving incorrect or poorly prepared food orders.

- 1868 – Unappetizing Presentation: Dishes that are poorly presented or lack visual appeal.
- 1869 – Unpleasant Ambience: Uncomfortable or uninviting atmosphere in the restaurant.
- 1870 – Dirty or Unhygienic Conditions: Lack of cleanliness in the dining area, restrooms, or
- 1871 utensils.
- 1872 – Limited Menu Options: A limited selection of dishes or lack of variety.
- 1873 – Poor Food Quality: Dishes that are poorly cooked, tasteless, or of low quality.
- 1874 – Overpriced: Excessive prices for the quality and portion sizes of the food.

1875 **I.8.2 Descriptive Details**

1876 We use cuisine types as the characteristics of restaurants, and the attributes are listed as follows:

- 1877 • Turkish
- 1878 • Spanish
- 1879 • Greek
- 1880 • Italian
- 1881 • French
- 1882 • American
- 1883 • Mexican
- 1884 • Canadian
- 1885 • Cajun
- 1886 • Tex-Mex
- 1887 • Brazilian
- 1888 • Peruvian
- 1889 • Argentinean
- 1890 • Colombian
- 1891 • Venezuelan
- 1892 • Ethiopian
- 1893 • Moroccan
- 1894 • South African
- 1895 • Nigerian
- 1896 • Egyptian
- 1897 • Chinese
- 1898 • Japanese
- 1899 • Indian
- 1900 • Thai
- 1901 • Korean
- 1902 • Australian
- 1903 • New Zealand
- 1904 • Polynesian
- 1905 • Hawaiian
- 1906 • Singaporean

1907 **J Examples for Filtered Attribute Values**

1908 Here we give some examples of the filtered attributes.

1909 For the Amazon product review dataset, some filtered attributes are listed as follows.

- 1910 • beauty:
 - 1911 – Hair Dryer (close to health and personal care)
 - 1912 – Hair Straightener (close to health and personal care)
- 1913 • electronics:
 - 1914 – Car dashcam (close to automotive)
 - 1915 – Wireless earbuds (close to cell_phones_service)
- 1916 • office_products:
 - 1917 – Mouse pad (close to electronics)

1918 For NYT dataset, some filtered attributes are listed as follows:

- 1919 • american_football:
 - 1920 – The economic impact of football on local communities and businesses.
 - 1921 – The role of nutrition and hydration in optimal football performance
 - 1922 – The history of football's integration with America's armed forces
- 1923 • stocks_and_bonds:
 - 1924 – The role of artificial intelligence in stock market prediction
 - 1925 – The effects of climate change on bond performance
- 1926 • movies:
 - 1927 – The future of cinema in a post-COVID world
 - 1928 – The role of international markets in Hollywood's global success

1929 For Reddit dataset, some filtered attributes are listed as follows:

- 1930 • amc_stock:
 - 1931 – Working in the stock market industry.
 - 1932 – Impact of COVID-19 on amc_stock.
 - 1933 – Having a background in finance or business.
- 1934 • what_car_should_i_buy:
 - 1935 – Towing capacity and trailer options
 - 1936 – Financing and leasing options
 - 1937 – Road trip and travel experiences
- 1938 • pittsburgh:
 - 1939 – Attending a game at PNC Park
 - 1940 – Watch a Steelers football game at Heinz Field

1941 For StackExchange dataset, some filtered attributes are listed as follows:

- 1942 • algorithm:
 - 1943 – Debugging an optimization algorithm for a complex logistics problem.
 - 1944 – Improving the robustness of an encryption algorithm against attacks.
 - 1945 – Finding a way to better handle overflow errors in a cryptographic algorithm.
- 1946 • iPhone:

- 1947 – Unable to use certain features in certain regions/countries
- 1948 – Trouble with GPS not working properly
- 1949 • opengl:
- 1950 – Difficulty setting up a custom camera in OpenGL
- 1951 – Inconsistent rendering of 3D models in OpenGL
- 1952 – Difficulty integrating OpenGL with a cross-platform game engine

1953 **K Examples for Merged Attribute on the arXiv dataset**

1954 Here we give some examples of the filtered attributes.

- 1955 • optimization techniques for computer vision:
 - 1956 – computer vision and pattern recognition
 - 1957 – machine learning
- 1958 • black holes:
 - 1959 – classical physics
 - 1960 – popular physics
- 1961 • evolving networks for financial trading:
 - 1962 – neural and evolutionary computing
 - 1963 – computational finance
- 1964 • multi-lingual information retrieval:
 - 1965 – information retrieval
 - 1966 – computation and language
- 1967 • iot operator systems and their security challenges:
 - 1968 – operating systems
 - 1969 – cryptography and security
- 1970 • complexity of social networks:
 - 1971 – computational complexity
 - 1972 – social and information networks

1973 **L The Prompt Format for MetaPrompt**

1974 In this section, we provide more details about the implementation of MetaPrompt [45]. It first lever-
 1975 ages LLMs to provide additional task-specific information, and then augments the class-dependent
 1976 prompt with the task-specific knowledge.

1977 **Prompts to ChatGPT:** [Original Prompt]. What does this task ask us to do?

1978 **Response from ChatGPT for NYT dataset:** The task of generating an example of a NYT news asks
 1979 the model to create high-quality, informative, and engaging news articles that mimic the writing style
 1980 and structure commonly found in The New York Times (NYT).

1981 **Response from ChatGPT for Amazon dataset:** The task of generating an example of a product
 1982 review asks the model to create a detailed and informative review for a particular product, highlighting
 1983 its features, performance, pros and cons, and overall user experience. This review should be helpful
 1984 for potential buyers, providing them with valuable insights to make an informed purchasing decision.

1985 **Response from ChatGPT for Reddit dataset:** The task of generating an example of a post similar
 1986 to Reddit asks the model to craft a well-written and engaging post that adheres to the guidelines and
 1987 community standards of the specific subreddit it’s intended for. The post should spark discussions,

1988 share valuable information, or showcase interesting content that resonates with the subreddit’s
1989 audience.

1990 **Response from ChatGPT for StackExchange dataset:** The task of generating an example of a post
1991 similar to StackExchange asks the model to compose a clear and concise question or answer related to
1992 a specific topic within the given StackExchange community. The post should be well-researched and
1993 supported by relevant sources or personal expertise, aiming to provide valuable insights or solutions
1994 to other users’ queries.

1995 **M Dataset Information**

1996 **M.1 Dataset Documentations**

1997 The dataset is provided in *json* format; there are three json files corresponding to the original *train*,
1998 *validation* and *test* split. We also include two additional file, *attrprompt* and *simprompt*,
1999 which is generated by AttrPrompt and SimPrompt as the synthetic training data.

2000 Each data point contains the following fields:

- 2001 • *label*: the label for the example. For multi-class classification, the *label* field is an
2002 integer, while for multi-label classification, the *label* field is a `list[int]` containing one
2003 or multiple integers as each example may refer to multiple classes;
- 2004 • *text*: a content of each example.

2005 **M.2 Intended Uses**

2006 AttrPrompt and SimPrompt are intended for researchers in machine learning, natural language
2007 processing, and related fields to innovate novel methods for training data generation problems.

2008 **M.3 Hosting and Maintenance Plan**

2009 The codebase is hosted and version-tracked via GitHub. It will be available under the link <https://github.com/yueyu1030/attrprompt>. The download link of all the datasets can be found in
2010 the Github repository.

2012 Note that it is a community-driven and open-source initiative. We are committed and have the
2013 resources to maintain and actively develop it for at minimum the next five years. We plan to grow the
2014 GitHub repo by including new tasks and datasets and warmly welcome external contributors.

2015 **M.4 Licensing**

2016 We license our work using Apache 2.0⁸. All of the *train/validation/test* data are publicly
2017 released by previous work [1, 32, 17].

2018 **M.5 Author Statement**

2019 We the authors will bear all responsibility in case of violation of rights.

⁸<https://www.apache.org/licenses/LICENSE-2.0>