Comparative Analysis of SLMs Performance with Various Preprocessing Techniques in Book Review Generation

Anonymous ACL submission

Abstract

Our project focuses on optimizing small language models (SLMs) by implementing specific preprocessing techniques and creating efficient datasets. This approach is designed to enhance the capabilities of models with less computational and memory cost, aiming to achieve results that are comparable to current SLMs. The emphasis is on refining data to its essential elements to improve grammatical accuracy, diversity, and basic reasoning skills in text generation. Our findings indicate that with targeted preprocessing and dataset optimization, SLMs can produce outcomes similar to those obtained with longer training times and larger datasets. We also report the successful training of an SLM on a carefully curated corpus of 0.96 million reviews within a span of 5 hours using a single V100 GPU, demonstrating our approach's efficiency in generating quality text completions.

1 Introduction

Language comprehension involves more than words, including grammar, syntax, and context. Recent advances in large language models (LLMs) show deep understanding of these aspects. Yet, the ability of smaller models like GPT-Neo and GPT-2 to mimic this understanding is still being researched. Despite their size, these models sometimes struggle with coherent sentences, likely due to language complexity and diverse training data

Our investigation focuses on the potential of training SLMs on customized datasets to enhance the quality of generated sentences. The goal is to reduce training costs for language models by creating optimized datasets for efficient training and effective modeling.

Our approach focuses on smaller, high-quality datasets to cut training time. We use low-cost preprocessing, like removing rare words, to reduce vocabulary size and simplify learning. This also involves discarding sentences with many uncommon words, and concentrating on learning common language patterns.

To assess small language models (SLMs), we used both automated and manual methods. GPT 3.5 served as an automated standard for assessing metrics like grammar, creativity, and consistency. Manual evaluations looked at causal and contrasting relations among other elements.

We hypothesized that datasets with sentences having at least one verb, noun, adjective, and adverb would be most effective. This structure aims to balance linguistic elements, helping SLMs create coherent, creative, and context-appropriate narratives.

2 Related Works

Our research on training smaller language models (SLMs) using curated datasets builds on and differs from previous work, specifically the study by Eldan and Li (Eldan and Li, 2023). While they employed GPT-3.5 for dataset construction, incurring higher costs, our method achieved comparable results more cost-effectively. Their training time was 24 hours, compared to just 5 hours for our models to be efficient.

Radford et al.'s (Radford et al., 2019) work on GPT models and Brown et al.'s (Brown et al., 2020) research on GPT-3 demonstrated the capabilities of larger models. Our study, however, focuses on achieving similar capabilities in SLMs with significantly reduced resource usage, particularly in narrative completion tasks.

The efficiency-focused research of Sanh et al. (Sanh et al., 2019) and Turc et al. (Turc et al., 2019) has informed our approach. Yet, our study is distinct in applying these principles to the narrative language domain, emphasizing the balance between model size, efficiency, and specific task performance.

3 Method

3.1 Reproducibility

We set zero temperature across all models mentioned in this report. The code is available at Training.ipynb and Manually Evaluation.ipynb. The models, training datasets, validation dataset are available on Huggingface named TinyReviews_raw/common/adv/adj, books_raw/common/adv/adj and kindle for reproducibility purpose.

3.2 Customized GPT-Neo Model

We use a model architecture similar to that designed by Eldan and Li (2023), which includes embedding layers with a hidden size of 64 and 8 transformer blocks. It also features a small Multi-Layer Perceptron (MLP) with 64 input features and 256 output features after the transformer layers, providing local information transformation at each position separately. Benefiting from its simplified layers containing 1 million parameters, this SLM is capable to generate coherent, grammatically correct sentences after only a few hours of training on a single V100 GPU.

3.3 Book Reviews Datasets

To explore the impact of different preprocessing techniques, we built a pipeline and constructed four datasets from the original Amazon Review data (Ni et al., 2019). Our pipeline consists of four steps:

1. The common preprocessing technique filters out reviews that are shorter than three words or longer than five hundred words. Super-short reviews do not improve the model's reasoning capability significantly, and super-long reviews may slow down the training speed.

2. Construct a common-word dictionary dataset with a vocabulary size 20,000 using 231,392 music instrument reviews. Although this dictionary contains some noise data, such as the actual uncommon words and words containing punctuations, it includes enough actual common words based on its 20,000 size. This approach, which does not involve computationally heavy preprocessing techniques such as lemmatization, still ensures that the processed reviews have a limited vocabulary size after our manual verification. Then, we filter out reviews that are outside the common words dictionary.

3. Utilize NLTK's POS tagger to retain reviews containing at least one verb, noun, and adjective.

4. Follow a strategy similar to the previous step but retain reviews containing at least one verb, noun, adverb, and adjective.

As shown in figure 1, the number of reviews decreases gradually after applying more preprocessing techniques, and even the smallest dataset takes 3-5 hours to train a tiny SLM on one V100 GPU. Another note is that each more extensive dataset includes all data in the smaller ones. From this pipeline, we construct a smaller dataset that has much higher quality.

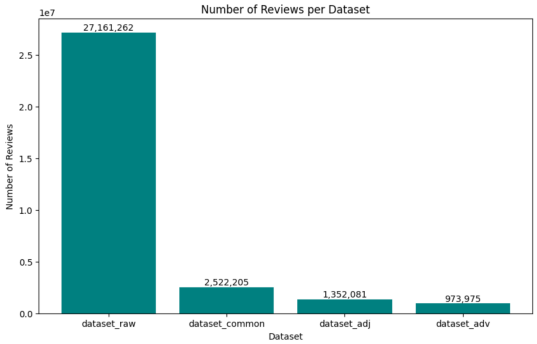
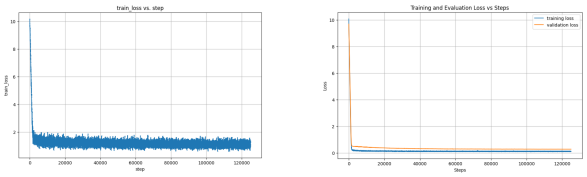


Figure 1: Number of reveiws in four customized datasets

3.4 Training process

We train four GPT-Neo models with the same architecture and randomly initialized weights on an equal amount of data, first 960,000 reviews from our four previously constructed datasets. Models are named as model_raw, model_common, model_adj and model_adv based on different training dataset. We only use one Kindle review in the validation dataset, which was not present in the book review training dataset to speed up the training speed. This setup allowed us to make a fair comparison among different preprocessing techniques, as they are the only variable between the four models. There are two interesting observations:



(a) Using raw dataset (b) Using filtered dataset

Figure 2: Training Loss comparison

1. Due to the significant amount of noise in the raw dataset, the training loss of the first model was

unstable, as shown in figure 2a. This indicates that the model spent time learning insignificant patterns.

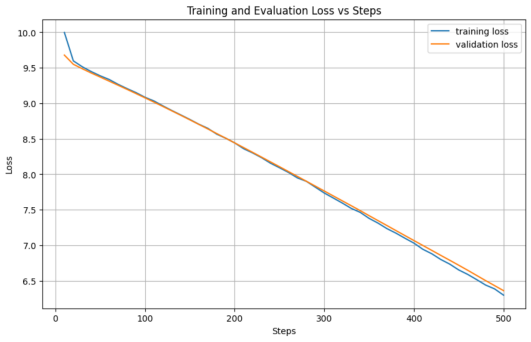


Figure 3: Validation loss VS Training loss

2. After approximately 200 steps (equivalent to 1600 reviews), the model consistently exhibited lower training loss compared to the validation loss in figure 3, suggesting that it begans to learn the common patterns from the training dataset in the very beginning.

4 Results

4.1 GPT-Eval: Utilize GPT-3.5 to evaluate the completions

We manually select 50 unseen reviews from kindle review validation set and save the first half of each review as test dataset. We also make sure each model can generate at least 5 words from each test senteces (first half part of reviews). Then we prompt GPT-3.5 to evaluate each completions on their grammar, creativity and consistency with the first half in a score range from 1 to 10.

Example interaction is (our models' generation is highlighted and GPT-3.5 generates the three scores in the end):

```
In the following exercise, student is given a beginning of a comment about Amazon Book Review. The student needs to complete it into a complete comment. The exercise tests the student's language abilities and creativity. The symbol *** marks the separator between the prescribed beginning and the student's completion:

I was hoping to find this one in book form. The story looks***like a good read. I would recommend it to anyone.

Now, grade the student's completion in terms of grammar, creativity, and consistency with the comment's beginning. Each metric is ranging from 0 to 10. Please provide the information as a number array only, without additional explanation or text.

[8, 7, 9]
```

4.1.1 Performance comparison between models

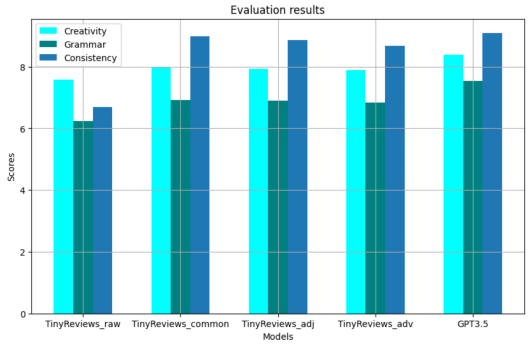


Figure 4: GPT-eval scores comparison

The models trained on the filtered dataset improve all three scores significantly compared to the one trained on the raw dataset, as shown in figure 4. The evaluation results from GPT-3.5 partially prove our hypothesis that higher-quality datasets lead to higher-quality completions from SLMs. The most significant improvement is observed in the grammar and consistency scores, indicating that the model trained on a filtered dataset is more capable of generating grammatically correct completions that also follow the logic and sentiment of the given prompts.

However, the differences in completion scores are subtle among the three models trained on different filtered datasets. One explanation is that these three straightforward metrics need to be more comprehensive, and we conducted a manual inspection later to evaluate the performance more carefully.

4.1.2 Performance comparison between checkpoints

After being trained on 40,000 of reviews, we observe that all models struggle to generate any meaningful content, shown on figure 5, although model_raw can sometimes provide a grammatical sentence. After being trained on half a million reviews, Model_adj and model_adv start to generate decent quality reviews. In contrast, model_common requires more data to produce comparable quality content, due to its training data has less knowledge about language structure.

	40000	480000	960000
model_raw	I have a book.	I was a good read and I was a lot of the characters...	I was a fan of the book...
model_common	I a book.	I was not like it. I was a good read.	I was not disappointed.
model_adj	I I I...	it was a good read.	I was very happy with it.
model_adv	I	it was a good read. I was a great read.	it was a good read. I enjoyed it very much.

Figure 5: Example completions after trained on different amount of reviews ("..." represents the repeated generations). We use the prompt, "I didn't read the book yet but", from the validation set.

More interestingly, model_raw consistently generates repetitive sentences even after being trained on 1 million data. This is likely because many unfiltered reviews are similar and short, which reduces the model's ability to capture the long dependencies among words.

4.2 Manually Evaluation using selected prompts

GPT-Eval works as a black-box analysis, so we manually use GPT-4 to conduct double evaluations using the same prompt. To ensure a thorough analysis, we include manual labeling due to the limited explainability of large language models.

4.3 Qualitative Analysis by GPT-4

We used a unique prompt, "I didn't finish the book yet but,". In Figure 6, evaluations using GPT-4 indicate that the SLM trained on raw data performed poorly, while the SLM trained with more structured POS data [Nouns, Verbs, Adjectives, Adverbs] (referred to as Adv) performed the best, followed by the common words model (Common). The SLM generated a comment demonstrating an understanding of the prompt's implied contradiction.

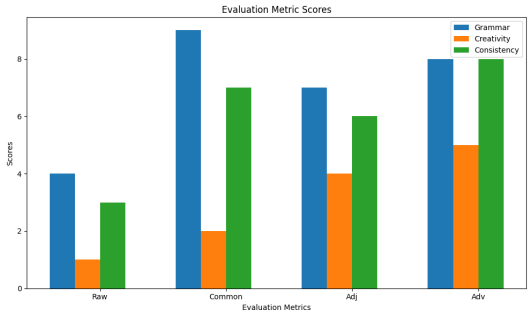


Figure 6: Advanced Evaluation using GPT-4

4.4 Context-Based Manual Labelling

We analyzed reviews based on everyday scenarios, assessing their reasoning and context. Comments with different beginnings were coded for clarity: Red for incorrect, Yellow for somewhat accurate but flawed, and Green for accurate responses. More in-depth analysis can be found in Manually Evaluation.ipynb.

4.4.1 Reviews with Causal Relationship

The prompt ended with a causal relationship (using "because"). Table 1 shows the Raw model's output was marked red for its incoherence and repetition. The Common and Adj models received yellow labels, partially addressing the prompt but missing its causal depth. The Common model veered off-topic, while the Adj model's reply was too superficial. In contrast, the Adv model earned a green label for its coherent, prompt-aligned response, demonstrating a clear grasp of the story's compelling elements.

Table 1: Analysis with Causal Relationship

Prompt: The story was captivating because
Raw: The story was captivating because the story was very well written. I was a bit of the characters and the characters were... (keep repetitive)
Common: The story was captivating because I was not sure that I was going to read it. I would recommend it to anyone.
Adj: The story was captivating because it was a little slow but it was a good read.
Adv: The story was captivating because I was a little more than I expected. I would recommend this book to anyone who is going to read.

4.4.2 Reviews with Contrasting Relationship

Table 2 indicates the Raw model received a yellow label for starting correctly but failing to offer meaningful contrast, instead repeating points. The Common, Adj, and Adv models earned green labels for effectively following the prompt's contrasting structure. The Common model subtly contrasts the book's uniqueness with enjoyment, the Adj model contrasts style recognition with reading uncertainty, and the Adv model contrasts style with story quality, maintaining context. These demonstrate a solid understanding of contrasting concepts.

4.4.3 Reviews with Comparative Statement

In Table 3, Raw model earned a red label for its nonsensical repetition, failing to distinguish the book's unique features. Similarly, the Common model also received red, offering no comparative analysis. In contrast, the Adj model's response was

Table 2: Analysis with Contrasting Relationship

Prompt: The author's style is unique, but
Raw: The author's style is unique, but it's a great read. I have read the book and I have read it.
Common: The author's style is unique, but it was a good read.
Adj: The author's style is unique, but I'm not sure I would like to read this book.
Adv: The author's style is unique, but the story is very good.

labeled yellow for its correct yet superficial comparison, recognizing the book's merits without specific contrasts. The Adv model stood out with a green label, effectively comparing the book to traditional novels by providing personal insights, indicating a superior grasp of comparative constructs. Overall, the Adv model proved most adept at generating coherent comments, showcasing stability and understanding of diverse sentence structures.

Table 3: Analysis with Comparative Relationship

Prompt: Unlike traditional novels, this book features
Raw: Unlike traditional novels, this book features a great story of a young woman ... (repetition occurs)
Common: Unlike traditional novels, this book features
Adj: Unlike traditional novels, this book features are great.
Adv: Unlike traditional novels, this book features is a great book. I have read it in a day. I am very happy with it.

5 Discussion

5.1 Small high qualilty dataset can be enough to generate coherent book reviews

We tested GPT-3.5 using the same 50 test reviews and found that our best model achieves performance comparable to GPT-3.5, shown on figure 4. Our GPT-neo model with 1M parameters trained on 1M sentences can generate coherent reviews after applying low-cost preprocessing techniques, such as removing uncommon words. In future work, readers can explore more low-cost techniques to improve dataset quality, which may further reduce the training costs for SLMs.

5.2 NLTK's POS taggers may introduce unfair advantages

In the data preprocessing pipeline, we utilize NLTK's POS tagger to construct dataset_adj and dataset_adv, then select an equal amount of data from each dataset aiming for a fair comparison. However, using the pretrained tagger introduces additional knowledge to model_adj and model_adv, since the tagger is trained on the Wall Street Journal corpus, which contains millions of words. We suggest only conducting algorithm-based preprocessing techniques in future work.

5.3 Models may benefit from longer sentence length

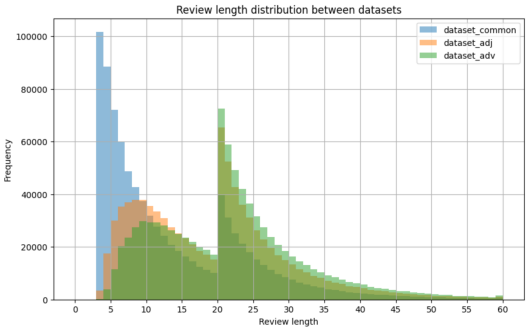


Figure 7: Review length distributions

Displayed in Figure 7, we find that most reviews from dataset_common contain fewer than 15 words, whereas reviews from dataset_adj and dataset_adv typically range from 20 to 30 words. Longer sentences may contain more knowledge than shorter ones, which in turn helps the model generate more coherent completions.

6 Conclusion

We presented a pipeline for generating datasets specifically tailored for SLMs. Our constructed dataset, where each sentence contains at least one verb, noun, adjective, and adverb, enabled our 1M parameter GPT-Neo to efficiently generate coherent book reviews in a shorter time. By incorporating a higher frequency of adverbs in training, the SLMs learned to create more emotionally expressive statements, achieving comparable results to LLMs. We hope that `TinyReviews` can inspire people to construct datasets from existing ones and build domain-specific GPTs on a limited budget.

7 Statement of contributions

Yang Kai Yam: Design SLM project flow. Conduct qualitative and manual analysis with coding and report.

Xu Michael: Construct datasets, train and evaluate the SLMs and write those in report.

Massoud Mahsa: Design SLM project flow. Conducting research through related works and evaluation of the overall project, writing the report.

References

Tom B. Brown et al. 2020. Language models are few-shot learners. In *NeurIPS*.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english?

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Alec Radford et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Victor Sanh et al. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Iulia Turc et al. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*.

Declaration
We acknowledge the use of GPT-4 to fix the grammar of the sentences we wrote at the final stage.