Forged-GAN-BERT: Authorship Attribution for LLM-Generated Forged Novels

Kanishka Silva
University of Wolverhampton, UK
a.k.silva@wlv.ac.uk

Ingo Frommholz
University of Wolverhampton, UK
ifrommholz@acm.org

Burcu Can
University of Stirling, UK
burcu.can@stir.ac.uk

Frédéric Blain
Tilburg University, NL
f.l.g.blain@tilburguniversity.edu

Raheem Sarwar
Manchester Metropolitan University, UK
r.sarwar@mmu.ac.uk

Laura Ugolini
University of Wolverhampton, UK
l.ugolini@wlv.ac.uk

Abstract

The advancement of generative Large Language Models (LLMs), capable of producing human-like texts, introduces challenges related to the authenticity of the text documents. This requires exploring potential forgery scenarios within the context of authorship attribution, especially in the literary domain. Particularly, two aspects of doubted authorship may arise in novels, as a novel may be imposed by a renowned author or include a copied writing style of a well-known novel. To address these concerns, we introduce Forged-GAN-BERT, a modified GAN-BERT-based model to improve the classification of forged novels in two data-augmentation aspects: via the Forged Novels Generator (i.e., ChatGPT) and the generator in GAN. Compared to other transformer-based models, the proposed Forged-GAN-BERT model demonstrates an improved performance with F1 scores of 0.97 and 0.71 for identifying forged novels in single-author and multi-author classification settings. Additionally, we explore different prompt categories for generating the forged novels to analyse the quality of the generated texts using different similarity distance measures, including ROUGE-1, Jaccard Similarity, Overlap Confident, and Cosine Similarity.

1 Introduction

Early applications of generative models for literary text generation go back to the works by Bailey (1974) for automatic poetry generation. Moreover, the most recent attempts to generate poems via text generative models were described by Saeed et al. (2019); Zhang and Lapata (2014); Yi et al. (2017); Wang et al. (2016); Yu et al. (2017); Liu et al. (2018); Beheitt and Hmida (2022). ChatGPT and other powerful generative models generated stories by investigating different prompting mechanisms (Benzon, 2023; Osone et al., 2021). In most recent attempts, the researchers have explored human-AI co-creation in literary areas, for instance, in the works of Calderwood et al. (2020); Frich et al. (2019). Also, the work in Uludag (2023) performed qualitative and quantitative methods to test the creativity of ChatGPT in psychology. Uludag (2023) finds that ChatGPT has some level of creativity but also imposes limitations, such as a limited understanding of the context and the inability to generate original ideas.

With the popularity of generative LLMs for creative content generation, there have been issues observed on well-known book-selling platforms such as Amazon, where AI-generated books are presented for sale under human writers' names with and without the original involvement of the authors (Friedman, 2023). Responding to this situation, platforms such as Amazon have taken measures, such as ordering self-publishing authors to explicitly declare whether their content is machine-generated (Radauskas, 2023). To address these challenges, organisations such as the 'Authors Guild' and 'The Society of Authors are actively pursuing legislative protection for human authors from such forged literary works under their names (Aut, 2023; SOA, 2023).

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 325–337

As a preliminary step to proposing possible solutions for such authorship issues, particularly considering a use case of machine-generated novels, we explored the ability to utilise GAN-BERT (Croce et al., 2020) to discriminate forged novels generated by ChatGPT from the texts of the original novels. The internal architecture of the GAN-BERT models combines a generator capable of generating fake texts similar to real ones. Since the GAN-BERT model already identifies fake texts (Silva et al., 2023), we want to test the hypothesis that it will perform well in detecting AI-generated novels in a similar style to the original novels. This paper presents the Forged-GAN-BERT model, specifically designed to identify forged novels within the context of authorship attribution. We utilised 20 novels per author during this study, considering 5 randomly selected authors, prompting ChatGPT to forge the books' styles with zero-shot prompting. In contrast to a recent study conducted by Jones et al. (2022), which is primarily on online posts, our research focuses on literary works. We utilise the GAN-BERT model to conduct a dual analysis of the forged texts, combining forged novels generated within the GAN generator and those created by LLM, like ChatGPT. Also, in Jones et al. (2022), they have used fine-tuning to generate AI text, but instead, we prompted ChatGPT to forge or disguise the author's style. To our knowledge, this is the first study using ChatGPT prompts to generate similar novelist styles and to utilise the GAN-BERT model to detect AI-generated novels.

Our study is steered by the following formulated research questions:

RQ 1 What are the implications of utilising various text similarity metrics in assessing the quality of forged novels?

RQ 2 Is it possible to distinguish between human novels vs LLM-generated novels with the Forged-GAN-BERT?

The remainder of the paper is organised into several sections: Section 2 provides a brief literature survey. Then, Section 3 describes the dataset information. Section 4 elaborates on the quality analysis of the forged novels against different prompt categories, emphasising the RQ1. Section 5 outlines the Forged-GAN-BERT model architecture related to the RQ2. Finally, Section 6 adds concluding remarks and future directions.

2 Related Work

Text generation models, aka Natural Language Generation (NLG), generate text closer or indistinguishable from human text or any other input format, such as image or video, which can be categorised into completion generation, text-to-text generation, and inference. Large Language Models (LLM) such as BERT (Devlin et al., 2019), T5 (Raffel et al., 2020), PALM Chowdhery et al. (2022), GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023b), were trained on larger datasets with billions of parameters, which can process massive loads of data and provide highly accurate results. BERT-based models (Devlin et al., 2019) and T5 (Raffel et al., 2020) were built on encoder-only or encoder-decoder architectures, respectively, and are flexible for adapting to many tasks by means of finetuning. Chowdhery et al. (2022) investigate the scalability factor of LLMs in terms of few-shot learning towards multilingual tasks and source code generation tasks. Recent LLM text generators mainly focus on the models' scalability and increasing the models' capacities compared to the predecessor models.

Advanced conversational models can be optimised for massive, high-quality data generation via prompt engineering (Saravia, 2022) on the API of interface level. By using prompt engineering in LLMs such as Flan (Chung et al., 2022), ChatGPT (OpenAI, 2023a), LLaMA (Hoffmann et al., 2022), and GPT-4 (OpenAI, 2023b), models can be utilised to curate new datasets (Wang et al., 2022; Sanh et al., 2022; Gehman et al., 2020; Bai et al., 2022) or as data augmentation strategies (Zhao et al., 2023; Shivagunde et al., 2023; Wang et al., 2023).

Mishra et al. (2022) discuss machine learning-based fake news detection techniques with a comparison to deep learning models. TweepFake (Fagni et al., 2021) detects DeepFake tweets generated by bots based on different text generation techniques such as RNN, Markov Chains, LSTM, and GPT-2. DeID-GPT (Liu et al., 2023) presents a zero-shot medical text de-identification based on GPT-4 in the domain of clinical notes.

In the area of authorship attribution, two main approaches exist for author identification: traditional approaches such as stylometric methods (Aborisade and Anwar, 2018; Soler Company and Wanner, 2017; Madigan et al., 2005), and deep learning-based approaches (Fabien et al., 2020;

Ruder et al., 2016; Saedi and Dras, 2021). Stylometric approaches focus on stylometric feature identification and utilising them in classification models. Moreover, ensemble models such as Bacciu et al. (2019); Moreau et al. (2015) combine stylometric and deep learning mechanisms to enhance the authorship attribution. Authorship Obfuscation, a sub-discipline of authorship attribution, specifically addresses hiding authors' writing styles and identifying such attempts (Dehouche, 2021; Jones et al., 2022).

The GAN-BERT model (Croce et al., 2020) integrates BERT-based models with the Semi-Supervised GANs, as illustrated in Figure 1. The GAN-BERT model is being used for a range of applications such as sentiment analysis (Myszewski et al., 2022; Ta et al., 2022), authorship attribution (Silva et al., 2023), text classification (Auti et al., 2022; Tanvir et al., 2022), and multi-task learning (Breazzano et al., 2021).

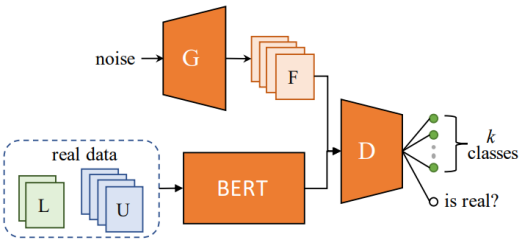

Figure 1: GAN-BERT Model (Croce et al., 2020)
G - Generator, D - Discriminator, F - Forged text, L - Labeled data, U - Unlabelled data

In contrast, considerable research has been performed on Fake News, Tweets, Medical, and Poems, but still limited attention to novels. Further analysing original and generated text, specifically for the literary domain, considering AI-generated forged text, has yet to be addressed.

3 Datasets

3.1 Original Novels

We used a subset of a 19th-century novelists' dataset created and curated from Project Gutenberg (Gutenberg) for the human-created texts. We selected 20 novels from 5 randomly selected authors: Arthur Conan Doyle, Henry Rider Haggard, Jack London, Mark Twain, and Wilkie Collins. The selected novel's list is in the released code repository [1].

[1]https://github.com/Kaniz92/Forged-GAN-BERT

3.2 Forged Novels:

In the literary domain, forgery can occur through two scenarios. One involves the misattribution of a text to a particular author(s), while the other involves copying a similar writing style. This writing style could be relevant to the author or the document itself. Our research focuses on the latter scenario, where we explore using LLMs to generate forged novels resembling existing original works and attempt to identify such creations.

Prompting the ChatGPT-3.5 API [2] has been used to generate similar novels per each original novel, ranging on different prompt categories: Length, Similarity, Identification, Chapter, and Temperature, illustrated in Table 1.

As explained in Table 2, the length parameter considers whether to include word count in the prompt query and an antecedent to the word count: 'at least', 'exactly', or 'at most'. The similarity parameter is defined to identify how ChatGPT interprets prompts to generate similar texts using antecedents to the book name:'similar to', 'as same as', 'same background as', and 'same characters as'. The identification parameter mentions the book text, i.e., with or without the author. There are different ways to prompt ChatGPT to generate novels, either to generate a full text or a chapter(s) explored in the Chapter parameter. In the ChatGPT API, the Temperature parameter can be set from 0 to 1, where a value closer to 1 generates creative texts. We used this dataset on different prompts to analyse the quality of the generated text but only utilised the Default prompt for the training and testing of the model. We prompted ChatGPT to forge the novel text in each prompt, not the author's style. All the prompts under each Prompt Category are mentioned in the Appendix A.

3.3 Preprocessing Datasets

As illustrated in Figure 2, the Project Gutenberg texts contain special header and footer sections. The Gutenberg sections were removed from the original dataset as a preliminary preprocessing step. Then, on both datasets, we performed typical preprocessing steps such as lowercasing, stopword removal, punctuation removal, and newline character removal. The cleaned original novel text has been prompted to the ChatGPT to generate forg-

[2]The forged novels were generated in March 2023. Hence, with the new ChatGPT-3.5 API update, the generated novels may differ from those used here.

Parameter Type	Prompt Example	Description
Length	Write a complete novel similar to {book_name} by {author}.	Without specifying a word limit
Similarity	Write a complete novel with same characters as {book_name} by {author}. The novel should be at least 10000 words.	Same fiction characters as the original novel
Identification	Write a complete novel similar to {book_name}. The novel should be at least 10000 words.	Without specifying the author
Chapter	Write the first chapter of a complete novel similar to {book_name} by {author}. The novel should be at least 10000 words.	First chapter only
Default	Write a complete novel similar to {book_name} by {author}. The novel should be at least 10000 words.	Default prompt with temperature set to 0.2

Table 1: Prompt examples per each parameter type. The temperature parameter is controlled via the ChatGPT parameters. The Default prompt was used to compare discriminative models.

Prompt Sub-Category	Prompt
Without	Write a complete novel similar to {book_name} by {author}.
Min	Write a complete novel similar to {book_name} by {author}. The novel should be at least 10000 words.
Exactly	Write a complete novel similar to {book_name} by {author}. The novel should be exactly 10000 words.
Max	Write a complete novel similar to {book_name} by {author}. The novel should be at most 10000 words.

Table 2: Prompt examples for Length Prompt Type. Other prompt examples can be referred in Appendix A.


```
\ufeffThe Project Gutenberg EBook of Little Novels, by Wilkie Collins\r\n\r\nThis eBook is for the use of
anyone anywhere at no cost and with\r\nalmost no restrictions whatsoever.  You may copy it, give it away
or\r\nre-use it under the terms of the Project Gutenberg License included\r\nwith this eBook or online at
www.gutenberg.org\r\n\r\n\r\nTitle: Little Novels\r\n\r\nAuthor: Wilkie Collins\r\n\r\nPosting Date: October
15, 2008 [EBook #1630]\r\nRelease Date: February, 1999\r\nLast Updated: September 11, 2016\r\n\r\nLanguage:
English\r\n\r\nCharacter set encoding: UTF-8\r\n\r\n*** START OF THIS PROJECT GUTENBERG EBOOK LITTLE NOVELS
***\r\n\r\n\r\n\r\n\r\n\r\nProduced by James Rusk\r\n\r\n\r\n\r\n\r\n\r\n\r\nLITTLE NOVELS\r\n\r\nBy Wilkie
Collins\r\n\r\n\r\n\r\n\r\n\r\nMRS. ZANT AND THE GHOST.\r\n\r\nI.\r\n\r\nTHE course of this narrative describes the
return of a disembodied\r\nspirit to earth, and leads the reader on new and strange ground.\r\n\r\nNot in the
obscurity of midnight, but in the searching light of day, did\r\nthe supernatural influence assert itself.
Neither revealed by a vision,\r\nnor announced by a voice, it reached mortal knowledge through the
sense\r\nwhich is least easily self-deceived: the sense that feels.
```

```
Title: The Mysterious Stranger\n\nChapter 1\n\nIt was a dark and stormy night when the stranger arrived in the
small village of Oakwood. He was a tall, thin man with piercing blue eyes and a sharp nose. He wore a long
black coat and a wide-brimmed hat that hid his face in shadows. The villagers watched him with suspicion as he
made his way to the local inn.\n\nThe innkeeper, Mr. Brown, greeted him with a forced smile. "Good evening,
sir. How may I assist you?"\n\nThe stranger looked at him with a cold stare. "I need a room for the
night."\n\n"Of course, sir. Right this way." Mr. Brown led him to a small room on the second floor. "Is there
anything else you need?"\n\n"No, thank you." The stranger closed the door behind him and locked it.\n\nMr.
Brown shook his head and went back to the bar. "Strange fellow," he muttered to himself.\n\nChapter 2\n\nThe
next morning, the stranger left the inn and walked to the village square. He watched the villagers going about
their daily business, studying their faces with interest. He seemed to be looking for someone.\n\nHe found her
in the local church. She was a young woman with long brown hair and green eyes.
```


Figure 2: Original Novel (top) and Forged Novel (bottom) - Little Novels by Wilkie Collins. These are raw texts from the Project Gutenberg and ChatGPT responses, respectively. Prior to the training and testing, several pre-processing steps are performed as in Section 3.

eries. While the features we excluded during the pre-processing stage are commonly employed as stylometric features in authorship studies, our focus for author classification with LLMs mainly involves text-based features.

4 Quality Analysis of Forged Novels

It is important to evaluate the quality of the generated forged novels (F) by comparing them with the original novels (O). Different prompt categories (P) have been considered, with $p \in$ {'Length', 'Similarity', 'Identification', 'Chapter', and 'Temperature'}. Although an infinite range of prompts can be used for these experiments, we considered only a finite set of 18 different prompts. Since the objective of each prompt is to generate a similar novel to a given original novel, we were interested in the generated text quality and the similarity, hence utilised a range of metrics such as ROUGE-1, Jaccard Similarity, Overlap Confident,

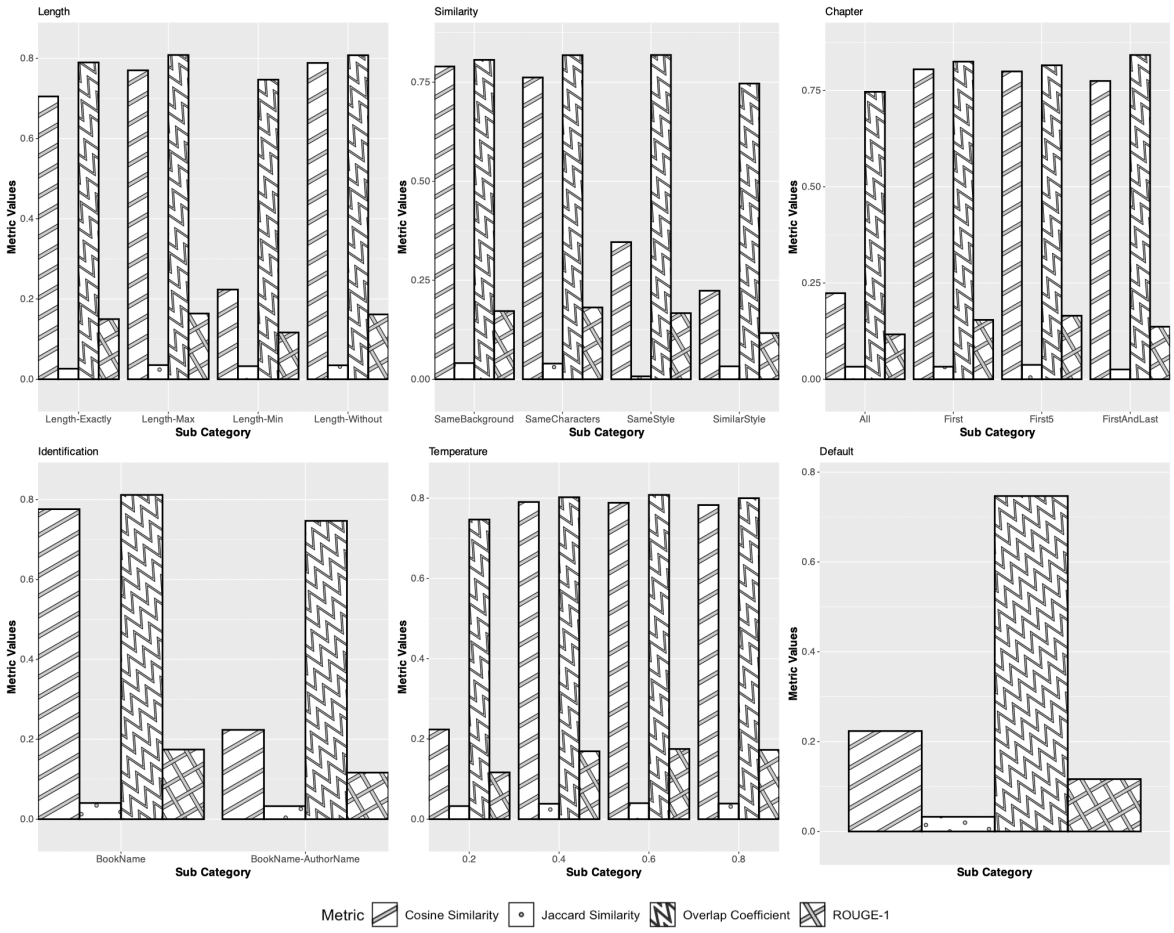

Figure 3: Prompt Type Impact Calculation using Similarity Scores

and Cosine Similarity. Each averaged similarity score can be calculated as follows, considering a basic averaging approach:

$$\text{Average Distance}_j^m = \frac{1}{N}\sum_{i=1}^{N} D^m_{(o_i,f_{ij})}$$

with o_i an original novel, p_j a prompt category and f_{ij} the corresponding forged novel of o_i generated using p_j. For a given similarity measure m, with $m \in \{$ROUGE-1, Jaccard Similarity, Overlap Coefficient, Cosine Similarity$\}$, the distance D between o_i and f_{ij} is represented as $D^m_{(o_i,f_{ij})}$. N is the total number of pairs of novels we average over.

For each prompt category P, the results of the averaged distribution for each prompt sub-category (see Table 2) are illustrated in Figure 3. These results indicate that the 'Chapter' prompt category has more impact on the generated text similarity based on the Overlap Coefficient and Cosine Similarity metrics. The 'Similarity' prompt category reports the highest ROUGE-1 score, which suggests that such prompts captured similarity better content-wise.

5 Forged-GAN-BERT Model

In the proposed model architecture as in Figure 4, we are considering two aspects in addressing forged texts in authorship attribution:

1. augmented novels via Forged Novels Generator

2. generated fake text via GAN-BERT

The proposed Forged-GAN-BERT model differs from the original GAN-BERT model (Croce et al., 2020) by incorporating a dual forged text analysis curated explicitly for the authorship attribution task. In contrast to the original model, only labelled data were used in this approach. Across different experiment settings, we provide the model with different ratios of forged novels and original novels and varying numbers of predicted classes via the discriminator (D). The original and forged novels

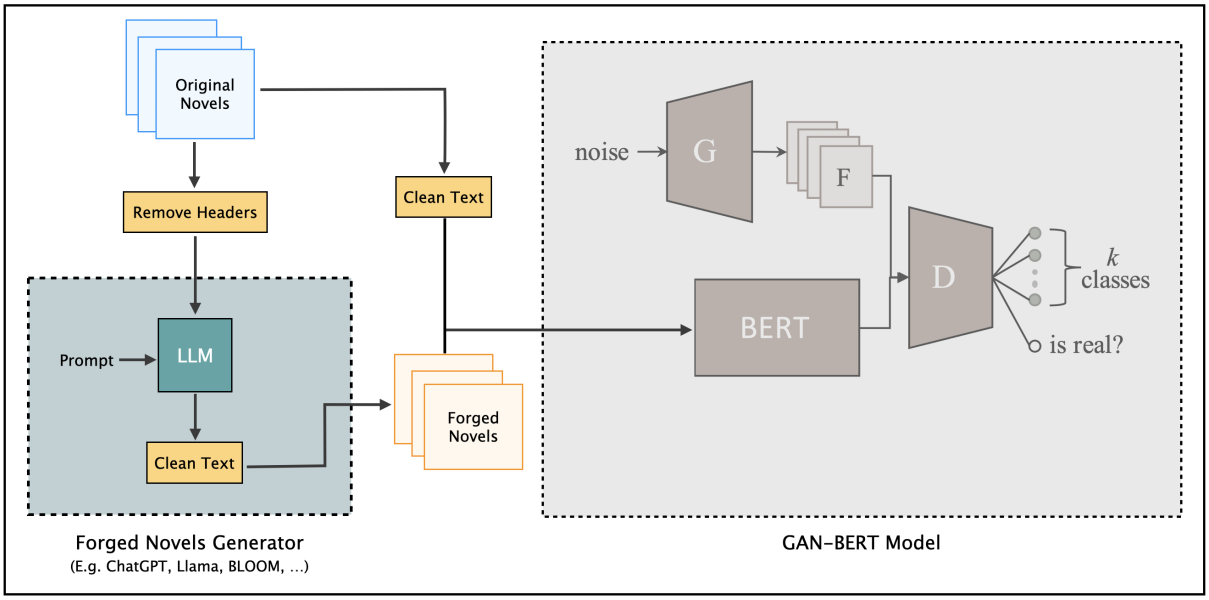

Figure 4: Forged-GAN-BERT Model Architecture. Original novels were given as context to the LLM-based prompting to generate Forged Novels, which concatenated as inputs to the BERT model (bert-base-cased) in the GAN-BERT model (Croce et al., 2020) (see Figure 1) to generate real text embedding. G and D represent the generator and discriminator, respectively. Calculated loss at D is used to update model parameters in both G and D. The forged novels generated by LLM and the fake text F from G contribute to the dual analysis of the forged novels.

are considered real data, passing through the BERT model to generate text embedding. Subsequently, using a controlled noise distribution from the latent space, the generator (G) produces fake text resembling the real text, which is used as an input to the discriminator. The calculated loss is then used to adjust the generator and discriminators' training parameters.

5.1 Dual Analysis of Forged Novels

Conventional approaches to detecting forged text typically combine generated or annotated forged text with the original text. In contrast, the proposed model performs a dual analysis by incorporating two types of forged texts: those derived from real-world sources and those generated by the GAN generator. This approach enhances the model's capability to detect forged novels, whether written by humans or machine-generated.

Furthermore, a secondary hypothesis examined via this dual analysis is that generating forged text based on existing forged text may reveal the imposter's true writing style. The fake novels F, generated for each forged novel, act as forged texts over existing forged texts.

5.2 Experiment Design

We designed the following experiments under different dataset slices generated from the Default prompt with the temperature value set to 0.2. Although the higher temperature values indicate more creativity in the generated text, we had to select a lower range value to ensure the generated text would keep the same style as the suggested novel.

1. ChatGPT as an Author Class

2. Human novels vs forged novels

When representing ChatGPT as an author class, we simulated a scenario of determining whether a test novel is a forgery against a known author's work. The comparison between human and forged novels evaluates the model's ability to identify LLM-generated texts across different authors, which evaluates the model's generalisation ability.

The BERT embeddings were used to represent the text, and the discriminator problem was modelled as simple text classification. The models were trained with default parameters wherever appropriate: a batch size of 8, 5 epochs, a warmup proportion of 0.1, a learning rate of 1e-5, a dropout rate of 0.2, and using Adam optimiser.

Model	F1	F1(Human)	F1(ChatGPT)	Accuracy	AUC
BERT	0.688 ±0.199	0.648	0.728	0.700	0.700
Longformer	**0.975 ±0.051**	**0.978**	0.971	**0.975**	**0.975**
RoBERTa	0.949±0.070	0.956	0.943	0.950	0.950
Forged-GAN-BERT	**0.975 ±0.057**	0.971	**0.978**	**0.975**	**0.975**

Table 3: Comparison between ChatGPT and All Authors (Averaged) Binary Classifications using BERT Embedding as features.

Model	F1	F1(Human)	F1(ChatGPT)	Accuracy	AUC
BERT	0.275	0.000	0.760	0.550	0.917
Longformer	0.389	0.100	**1.000**	0.675	**1.000**
RoBERTa	0.397	0.080	**1.000**	0.700	**1.000**
Forged-GAN-BERT	**0.710**	**0.600**	**1.000**	**0.850**	**1.000**

Table 4: Comparison between ChatGPT vs Human Binary Classifications using BERT Embedding as features.

5.3 ChatGPT as an Author Class

At the primitive level, we investigated the model performance when ChatGPT-forged novels were compared to a single author based on binary classification. We trained author-based models with 20 novels from the original author and 20 ChatGPT forgeries for each novel, resulting in a balanced uniform dataset slice. We averaged results obtained per author to obtain a better generalisation.

The classification against a single author was performed by reporting Accuracy, F1, AUC scores, and each class F1 as illustrated in Table 3. The dataset was well balanced during each author's comparisons, with 20 novels from the author and 20 ChatGPT novels per each, resulting in 40.

The high accuracy of 0.98 and F1 of 0.97 indicate a superior performance of Forged-GAN-BERT in distinguishing forged novels and each original novel. For instance, consider a scenario where a bookseller would want to investigate whether a specific novel is a forgery based on a known author's work. With a higher number of works to compare in real life, manual processing becomes impractical and time-consuming. Instead, the proposed model suggests an automated process that can be integrated into such a scenario.

The AUC of 0.97 indicates the dataset balance between the two classes. F1(Human) and F1(ChatGPT) scores evaluated the model performance if only a particular class is present in the dataset; for example, if only authors' original work is presented to the model, then it is capable of identifying correct authors with a 0.97 of F1 score, and with 0.98 of F1 score vice versa.

Compared to the baseline models, BERT shows a lower accuracy of 0.70 and F1 of 0.69, suggesting a slightly weaker performance.

5.4 Human novels vs forged novels

To experiment with the model performance in the multiple authors' scenario, we have mixed all the ChatGPT forgeries with original novels. We used the 100 original novels from our 5 human novelists and their forged counterparts generated by ChatGPT, resulting in a balanced distribution. The stratified k-fold sampling was used to overcome the overfitting. We performed another set of binary-class experiments using the same dataset by grouping all authors into the 'human' class and keeping the ChatGPT class the same.

In the multiple-author scenario, we considered ChatGPT as a unique author with five other authors, resulting in 6 classes to discriminate. We used the same models and parameter settings to experiment on this and reported the same metrics as in Table 4. For AUC, we used one-vs-rest in a multi-class setting, using ChatGPT class. The dataset is imbalanced in class distribution as ChatGPT text is five times each author's novel count, but it was balanced regarding the human vs AI text ratio.

The Forged-GAN-BERT model achieved a high accuracy of 0.85 and an F1 of 0.71, showing its ability to collectively identify human- and machine-generated novels. The perfect AUC score suggests a perfect separation between the two classes.

Compared to the baseline BERT model, which exhibits lower accuracy 0.55 and F1 0.25, indicates a weaker performance when distinguishing between human and ChatGPT-generated novels than the Forged-GAN-BERT model.

5.5 Robustness of the model

The Table 3 results were obtained by comparing different models per each author and getting the aver-

age of all the results. As per the standard deviation results recorded, it shows that both Forged-GAN-BERT and Longformer shows comparatively lower standard deviation across different authors, hence, both are robust over different author-wise comparisons. Although the Longformer model shows a competitive performance with the Forged-GAN-BERT model, it does not consist of a component to generate fake texts or to implicitly compare fake text vs real text.

Other models, BERT and RoBERTa are not comparatively successful in this case. Specifically, when comparing single-authors and multiple authors, the BERT model significantly showcases the lowest performance across almost all the metrics for both cases. This shows that the BERT models are not recommended for classifying forged novels, compared to the other models.

Further, the Table 4 results were obtained by observing one model to compare human vs ChatGPT novels, where Forged-GAN-BERT outperforms all other models across all the metrics. Altogether, we can deduce that the proposed Forged-GAN-BERT model is equally capable of identifying forged novels per each author or with multiple authors.

6 Conclusion

In conclusion, the introduced Forged-GAN-BERT model addresses the challenges of authorship attribution in machine-generated forged novels, explicitly and implicitly considering a dual forged text analysis approach. The results suggest that the proposed model outperforms other considered baseline models in identifying forged novels in single-author and multi-author classifications. Additional evaluation on the generated forged novels against different prompts utilised various similarity distance metrics such as ROUGE-1, Jaccard Similarity, Overlap Coefficient, and Cosine Similarity. The reported results indicated that the 'Chapter' configurations have more impact on generating novels similar to the original text. This evaluation can be extended for a probabilistic distribution approach to evaluate the forged novels for all the possible prompts in the infinite series of the prompts.

Future Work

We suggest that more research should focus on a proper evaluation mechanism of the similarity measure for literary works such as novels. Future directions could be around the authorship attribution area, focusing on stylistic-related features. Further, comparing and adhering to authorship obfuscation techniques would be an interesting future direction.

Although we utilised existing metrics, further research may be needed to evaluate the similarity between original and generated novels using language models such as ChatGPT, specifically on the creative index aspects.

This calculation can be extended considering a discrete probability distribution approach to determine the overall error rate, which suits future investigations. Further, integrating stylometric features into such probabilistic distribution would be another exciting direction.

Limitations

While this study unveils valuable insights into using the Forged-GAN-BERT model for authorship attribution in the context of forged novel scenarios, there are a few limitations to acknowledge. We only focused on 5 authors and 20 novels from each in a controlled dataset setting, denoting a close-set authorship attribution. In a real-world setting, we cannot expect the model to evaluate a text that may be a forgery of known classes; hence, further works should be investigated upon open-set authorship to ensure a more generalisation. As per the copyright considerations and issues with releasing forged novels, we refrain from releasing the entire dataset; instead, we have provided guidelines to reproduce the experiment settings.

Further, we acknowledge the character limitations imposed by the ChatGPT-3.5 model, which generates the forged novels, resulting in segments of the novels closely resembling the originals. To ensure consistency, we maintained the same text lengths as the original and generated forged novels during the experiments.

Ethics Statement

The selected original novels from Project Gutenberg (Gutenberg) between 1800 and 1914, out of the copyright duration as mentioned in 'Rule 1: Works First Published Before 95 Years Ago and Before 1977' and 'Rule 10(c) - Works of Treaty Parties and Proclamation Countries First Published Between 1923 and 1977'. Yet, we are not releasing the datasets to the public to prevent any unethical usage of the generated forged novels. The text generated in the generator in the Forged-GAN-BERT model is not human-readable; instead, it embeds

representations, preventing unethical usage. Any extended applications of this research should adhere to established ethical guidelines, such as using the generated forged novels and the proposed model only for classification purposes and research objectives. Moreover, using the proposed model and dataset generation should refrain from distributing any author's original content without appropriate consent.

Acknowledgements

This work is supported by the RIF-4 RIGHT project funded by the University of Wolverhampton, United Kingdom.

References

2023. Artificial Intelligence. https://authorsguild.org/advocacy/artificial-intelligence/. Accessed: 2023-10-30.

2023. Artificial Intelligence. https://www2.societyofauthors.org/where-we-stand/artificial-intelligence/#:~:text=Creators%20must%20be%20asked%20before,review%20by%20a%20human%20assessor. Accessed: 2023-10-30.

Opeyemi Aborisade and Mohd Anwar. 2018. Classification for Authorship of Tweets by Comparing Logistic Regression and Naive Bayes Classifiers. In *2018 IEEE International Conference on Information Reuse and Integration, IRI 2018*, pages 269–276. IEEE.

Tapan Auti, Rajdeep Sarkar, Bernardo Stearns, Atul Kr. Ojha, Arindam Paul, Michaela Comerford, Jay Megaro, John Mariano, Vall Herard, and John P. McCrae. 2022. Towards Classification of Legal Pharmaceutical Text using GAN-BERT. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 52–57, Marseille, France. European Language Resources Association.

Andrea Bacciu, Massimo La Morgia, Alessandro Mei, Eugenio Nerio Nemmi, Valerio Neri, and Julinda Stefa. 2019. Cross-Domain Authorship Attribution Combining Instance Based and Profile-Based Features. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019*, volume 2380 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *CoRR*, abs/2204.05862.

Richard W Bailey. 1974. Computer-assisted poetry: the writing machine is for everybody. *Computers in the Humanities*, pages 283–295.

Mohamed El Ghaly Beheitt and Moez Ben Haj Hmida. 2022. Automatic Arabic Poem Generation with GPT-2. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence, ICAART 2022, Volume 2, Online Streaming, February 3-5, 2022*, pages 366–374. SCITEPRESS.

William Benzon. 2023. Chatgpt Tells Stories, and a Note about Reverse Engineering: a Working Paper. *SSRN Electronic Journal*.

Claudia Breazzano, Danilo Croce, and Roberto Basili. 2021. MT-GAN-BERT: Multi-Task and Generative Adversarial Learning for Sustainable Language Processing. In *Proceedings of the Fifth Workshop on Natural Language for Artificial Intelligence (NL4AI 2021) co-located with 20th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2021), Online event, November 29, 2021*, volume 3015 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B. Chilton. 2020. How Novelists Use Generative Language Models: An Exploratory User Study. In *Joint Proceedings of the Workshops on Human-AI Co-Creation with Generative Models and User-Aware Conversational Agents co-located with 25th International Conference on Intelligent User Interfaces (IUI 2020), Cagliari, Italy, March 17, 2020*, volume 2848 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek

Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *CoRR*, abs/2204.02311.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *CoRR*, abs/2210.11416.

Danilo Croce, Giuseppe Castellucci, and Roberto Basili. 2020. GAN-BERT: Generative Adversarial Learning for Robust Text Classification with a Bunch of Labeled Examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2114–2119. Association for Computational Linguistics.

Nassim Dehouche. 2021. Plagiarism in the Age of Massive Generative Pre-Trained Transformers (Gpt-3). *Ethics in Science and Environmental Politics*, 21:17–23.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Maël Fabien, Esaú Villatoro-Tello, Petr Motlícek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for Authorship Attribution. In *Proceedings of the 17th International Conference on Natural Language Processing, ICON 2020, Indian Institute of Technology Patna, Patna, India, December 18-21, 2020*, pages 127–137. NLP Association of India (NLPAI).

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. TweepFake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415.

Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, Michael Mose Biskjaer, and Peter Dalsgaard. 2019. Mapping the Landscape of Creativity Support Tools in HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 389. ACM.

Jane Friedman. 2023. I Would Rather See My Books Get pirated than This (or: Why Goodreads and Amazon Are Becoming Dumpster Fires).

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 3356–3369. Association for Computational Linguistics.

Project Gutenberg. Project Gutenberg.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training Compute-Optimal Large Language Models. *CoRR*, abs/2203.15556.

Keenan Jones, Jason R. C. Nurse, and Shujun Li. 2022. Are You Robert or RoBERTa? Deceiving Online Authorship Attribution Models Using Neural Text Generators. In *Proceedings of the Sixteenth International AAAI Conference on Web and Social Media, ICWSM 2022, Atlanta, Georgia, USA, June 6-9, 2022*, pages 429–440. AAAI Press.

Dayiheng Liu, Quan Guo, Wubo Li, and Jiancheng Lv. 2018. A Multi-Modal Chinese Poetry Generation Model. In *2018 International Joint Conference on Neural Networks, IJCNN 2018, Rio de Janeiro, Brazil, July 8-13, 2018*, pages 1–8. IEEE.

Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, Tianming Liu, Dajiang Zhu, and Xiang Li. 2023. DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4. *CoRR*, abs/2303.11032.

David Madigan, Alexander Genkin, David D. Lewis, Er Genkin David D. Lewis, Shlomo Argamon, Dmitriy Fradkin, Li Ye, and David D. Lewis Consulting. 2005. Author Identification on the Large Scale. In *In Proc. of the Meeting of the Classification Society of North America*.

Shubha Mishra, Piyush Shukla, and Ratish Agarwal. 2022. Analyzing Machine Learning Enabled Fake News Detection Techniques for Diversified datasets. *Wireless Communications and Mobile Computing*, 2022:1–18.

Erwan Moreau, Arun Kumar Jayapal, Gerard Lynch, and Carl Vogel. 2015. Author Verification: Basic Stacked Generalization Applied To Predictions from a Set of Heterogeneous Learners - Notebook for PAN at CLEF 2015. In *CLEF*.

Joshua J. Myszewski, Emily Klossowski, Patrick Meyer, Kristin Bevil, Lisa Klesius, and Kristopher M. Schroeder. 2022. Validating GAN-BioBERT: A Methodology for Assessing Reporting Trends in Clinical Trials. *Frontiers Digit. Health*, 4:878369.

OpenAI. 2023a. ChatGPT - OpenAI Blog. `https://openai.com/blog/chatgpt/`.

OpenAI. 2023b. GPT-4 Technical Report. *CoRR*, abs/2303.08774.

Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. 2021. BunCho: AI Supported Story Co-Creation via Unsupervised Multitask Learning to Increase Writers' Creativity in Japanese. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama Japan, May 8-13, 2021, Extended Abstracts*, pages 19:1–19:10. ACM.

Gintaras Radauskas. 2023. Amazon orders self-publishers to disclose AI-generated content - cybernews.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Sebastian Ruder, Parsa Ghaffari, and John G. Breslin. 2016. Character-level and Multi-channel Convolutional Neural Networks for Large-scale Authorship Attribution. *CoRR*, abs/1609.06686.

Chakaveh Saedi and Mark Dras. 2021. Siamese networks for large-scale author identification. *Comput. Speech Lang.*, 70:101241.

Asir Saeed, Suzana Ilic, and Eva Zangerle. 2019. Creative GANs for generating poems, lyrics, and metaphors. *CoRR*, abs/1909.09534.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Elvis Saravia. 2022. Prompt Engineering Guide. *https://github.com/dair-ai/Prompt-Engineering-Guide*.

Namrata Shivagunde, Vladislav Lialin, and Anna Rumshisky. 2023. Larger probes tell a different story: Extending psycholinguistic datasets via in-context learning. *arXiv preprint arXiv:2303.16445*.

Kanishka Silva, Burcu Can, Frédéric Blain, Raheem Sarwar, Laura Ugolini, and Ruslan Mitkov. 2023. Authorship Attribution of Late 19th Century Novels using GAN-BERT. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 310–320. Association for Computational Linguistics.

Juan Soler Company and Leo Wanner. 2017. On the Relevance of Syntactic and Discourse Features for Author Profiling and Identification. In *Proceedings EACL 2017*, pages 681–687. Association for Computational Linguistics.

Hoang Thang Ta, Abu Bakar Siddiqur Rahman, Lotfollah Najjar, and Alexander F. Gelbukh. 2022. GAN-BERT: Adversarial Learning for Detection of Aggressive and Violent Incidents from Social Media. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022), A Coruña, Spain, September 20, 2022*, volume 3202 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Raihan Tanvir, Md. Tanvir Rouf Shawon, Md. Humaion Kabir Mehedi, Md Motahar Mahtab, and Annajiat Alim Rasel. 2022. A GAN-BERT Based Approach for Bengali Text Classification with a Few Labeled Examples. In *Distributed Computing and Artificial Intelligence, 19th International Conference, DCAI 2022, L'Aquila, Italy, 13-15 July 2022*, volume 583 of *Lecture Notes in Networks and Systems*, pages 20–30. Springer.

Kadir Uludag. 2023. Testing Creativity of ChatGPT in Psychology: Interview with ChatGPT. *SSRN Electronic Journal*.

Congcong Wang, Gonzalo Fiz Pontiveros, Steven Derby, and Tri Kurniawan Wijaya. 2023. STA: Self-controlled Text Augmentation for Improving Text Classifications. *CoRR*, abs/2302.12784.

Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. 2016. Chinese Poetry Generation with Planning based Neural Network. In *COLING 2016, 26th International Conference on*

Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan, pages 1051–1060. ACL.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2022. DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models. *CoRR*, abs/2210.14896.

Xiaoyuan Yi, Ruoyu Li, and Maosong Sun. 2017. Generating Chinese Classical Poems with RNN Encoder-Decoder. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data - 16th China National Conference, CCL 2017, - and - 5th International Symposium, NLP-NABD 2017, Nanjing, China, October 13-15, 2017, Proceedings*, volume 10565 of *Lecture Notes in Computer Science*, pages 211–223. Springer.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 2852–2858. AAAI Press.

Xingxing Zhang and Mirella Lapata. 2014. Chinese Poetry Generation with Recurrent Neural Networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 670–680. ACL.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq R. Joty. 2023. Retrieving Multimodal Information for Augmented Generation: A Survey. *CoRR*, abs/2303.10868.

A Appendix - Prompt Examples

Prompt Category	Prompt Sub-Category	Prompt
Length	Without	Write a complete novel similar to {book_name} by {author}.
	Min	Write a complete novel similar to {book_name} by {author}. The novel should be at least 10000 words.
	Exactly	Write a complete novel similar to {book_name} by {author}. The novel should be exactly 10000 words.
	Max	Write a complete novel similar to {book_name} by {author}. The novel should be at most 10000 words.
Similarity	SimilarStyle	Write a complete novel similar to {book_name} by {author}. The novel should be at least 10000 words.
	SameStyle	Write a complete novel as same as {book_name} by {author}. The novel should be at least 10000 words.
	SameBackground	Write a complete novel with same background in {book_name} by {author}. The novel should be at least 10000 words.
	SameCharacters	Write a complete novel with same characters in {book_name} by {author}. The novel should be at least 10000 words.
Identification	BookName	Write a complete novel similar to {book_name}. The novel should be at least 10000 words.
	BookNameAuthorName	Write a complete novel similar to {book_name} by {author}. The novel should be at least 10000 words.
Chapter	FirstAndLast	Write the first and last chapters of a novel similar to {book_name} by {author}. The novel should be at least 10000 words.
	All	Write a complete novel similar to {book_name} by {author}. The novel should be at least 10000 words.
	First	Write the first chapter of a complete novel similar to {book_name} by {author}. The novel should be at least 10000 words.
	First5	Write first five chapters of a complete novel similar to {book_name} by {author}. The novel should be at least 10000 words.
Default		Write a complete novel similar to {book_name} by {author}. The novel should be at least 10000 words.

Table 5: Prompts per each parameter type: The temperature parameter is controlled via the ChatGPT parameters for the Default prompt.