Fiction-Writing Mode: An Effective Control for Human-Machine Collaborative Writing

Wenjie Zhong[1,3], **Jason Naradowsky**[1], **Hiroya Takamura**[3]
Ichiro Kobayashi[2,3], **Yusuke Miyao**[1,3]
[1]The University of Tokyo, [2]Ochanomizu University,
[3]National Institute of Advanced Industrial Science and Technology
{zvengin, narad, yusuke}@is.s.u-tokyo.ac.jp
takamura.hiroya@aist.go.jp, koba@is.ocha.ac.jp

Abstract

We explore the idea of incorporating concepts from writing skills curricula into human-machine collaborative writing scenarios, focusing on adding *writing modes* as a control for text generation models. Using crowd-sourced workers, we annotate a corpus of narrative text paragraphs with writing mode labels. Classifiers trained on this data achieve an average accuracy of $\sim 87\%$ on held-out data. We fine-tune a set of large language models to condition on writing mode labels, and show that the generated text is recognized as belonging to the specified mode with high accuracy.

To study the ability of writing modes to provide fine-grained control over generated text, we devise a novel turn-based text reconstruction game to evaluate the difference between the generated text and the author's intention. We show that authors prefer text suggestions made by writing mode-controlled models on average 61.1% of the time, with satisfaction scores 0.5 higher on a 5-point ordinal scale. When evaluated by humans, stories generated via collaboration with writing mode-controlled models achieve high similarity with the professionally written target story. We conclude by identifying the most common mistakes found in the generated stories. The datasets and codes are available at the Github[1].

1 Introduction

Large-scale pre-trained language models (PLMs) have demonstrated a remarkable aptitude for generating text with an exceptional degree of fluency and structure (Guan et al., 2021; Tan et al., 2021), sparking renewed efforts to utilize them for the purpose of generating narrative fiction. Recent work has explored various ways of controlling PLMs, using sentiment (Luo et al., 2019), style (Kong et al., 2021a), and even character information (Liu et al., 2020a), in an attempt to cater the generated text to an author's intentions.

[1]https://github.com/ZVengin/fride

Summary: A boy was walking along a path in the forest, when he came across a heap of leaves.

Mode	Story
Dialogue	"Here are some leaves," he **whispered**. "They were wet when we came, and are wet now. I'll lie them down and wait." "What is it?" I **exclaimed**. "If you will stand still," **said** my boy, "I will show you … "
Action	He **stood** for a moment **looking** at me, then quietly he **picked up** the leaves, and **carrying** them in his hand, **climbed** to the top of the heap, and **examined** them …
Description	This heap consisted of **dead leaves**, many of them **rotten**, and still **wet**, with one or two **lying flat** on the ground, others **lying up against** the branches. The first to fall was the one I had thought **dead**. It had been crushed by the wind. …

Figure 1: Example of expanding the **Summary** into stories using different writing **Mode**s. The bold words imply the particular manner of expression in that mode. **Dialogue** focuses on the utterances spoken by characters, **Action** on the motion of characters, and **Description** on the depiction of characters or places.

However, the aforementioned controls deal primarily with *static* text attributes; an attribute like style is more synonymous with an entire author or book than with a single passage of text. Less attention has been paid to designing effective control factors for the real demands of human authors in collaborative writing settings, where authors typically exercise more *dynamic* control over their writing, varying attributes of the text at the sentence or paragraph level. Here we find inspiration from the creative writing literature, where the notion of a *fiction writing mode* is often presented as an important concept to consider when crafting narrative fiction (Klaassen, 2015).

A fiction-writing mode (also referred to as a rhetorical mode) is a particular manner of writ-

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1752–1765
May 2-6, 2023

Dia.	Act.	Des.	Unc.	Total	Len.(std)	Kappa
370	385	300	681	1,736	110(52)	0.64

Table 1: The number of instances for dialogue (**Dia.**), action (**Act.**), description (**Des.**), and uncertain (**Unc.**) modes in the dataset. **Kappa** is the inter-annotator agreement and **Len.(std)** is the average token number in each instance and its standard deviation.

ing, encapsulating the focus, style, and pacing of the text (among other things). Figure 1 illustrates how the same event can be described in different ways depending on the writing mode, using the three most common types, *Dialogue*, *Action*, and *Description*. Skilled authors proficiently use writing modes as a stylistic choice to engage readers and progress the narrative (see Section 2 for more detail) (Klaassen, 2015). Thus, we hypothesize that conditioning text generation models on writing modes can provide important controls to authors in a human-machine collaborative writing scenario.

To verify this hypothesis, we are faced with two challenges. First, to the best of our knowledge, there is no available dataset annotated with writing modes to train generation models. We create a Fiction wRIting moDE dataset (**FRIDE** dataset) containing 1,736 fiction paragraphs annotated by crowd-source workers with the three writing mode labels. Subsequently, we train a classifier on the **FRIDE** dataset and use it to annotate paragraphs of a large fiction corpus in order to create a larger-scale dataset. Using the established paradigm of training conditional text generation models by summarizing and reconstructing text (Sun et al., 2020), the dataset is used to train models which can be conditioned on a writing mode label.

Second, to measure whether writing modes allow for more effective control of text generation models, we need to compare the generated texts to the author's intention, which is unobservable. We design a new evaluation framework for human-machine collaborative writing where the author is given a paragraph of text, and is asked to recreate it solely through interaction with generative models. By using the paragraph as a proxy for the author's intention, we are able to assess the similarities between the intention and the generated text, and analyze the differences as indications of where current controls fail.

Through both automatic and human evaluation we show: (1) the use of writing mode labels with conditional text generation models contributes to average 1.4 and 2.0 points improvements on ROUGE-L and BERTScore; (2) the writing modes of generated text are effectively controlled, and are classified as belonging to the target mode in 87.6% of cases; (3) authors are 22.2% more likely to choose the suggestions from writing mode-controlled models, and assign them an average 0.5 higher satisfaction score (on an ordinal 1-5 scale) compared to the uncontrolled models; (4) applying writing mode control to collaborative writing enhances the similarities between the generated text and the authors' intention.

2 Fiction-Writing Mode Dataset

Fiction-writing modes have long been proposed as a useful abstraction in the study of literature and creative writing (Morrell, 2006; Klaassen, 2015), dating as far back as Aristotle (Halliwell and Aristotle, 1998). While there is no consensus on the categorization of writing modes, most sources prefer to introduce at least three modes: (1) *Dialogue*, direct quotation of characters speaking, (2) *Action*, an account of a series of events, one after another, chronologically, and (3) *Description*, a more detailed inspection of people, places, or things and their properties. These are the three major writing modes which are the focus of study in this paper.

Just as there is no agreement on how best to categorize writing modes, there is also no consensus on what text exhibits a particular mode. Even a single sentence can exhibit multiple writing modes, in varying degrees. However, for the purpose of this work, we assume that each paragraph can be categorized as exhibiting a single writing mode.

FRIDE Dataset In order to train models which generate text in a specified writing mode, we must first create a dataset, which we refer to as Fiction-wRIting moDE dataset (**FRIDE** dataset), which pairs paragraphs of narrative text with their corresponding writing mode labels. However, directly annotating writing modes on a large-scale narrative dataset is expensive and time-consuming. We first collect a modestly sized dataset from crowd-sourced workers, and utilize it to train a writing mode classifier. The classifier can then be used to provide high confidence labels to a much larger dataset of narrative text paragraphs, on a scale suitable for training large text generation models.

Paragraphs for annotation are collected from fic-

	Precision	Recall	F1
BERT-base	85.7	85.0	85.0
XLNet-base	84.7	84.4	84.3
RoBERTa-base	**86.3**	**85.2**	**85.2**

Table 2: The performance of writing mode classifiers on the **FRIDE** dataset.

tion books sourced from Project Gutenberg[2] (128 books) and, for more contemporary writing, Smashwords[3] (150 books). Each book is divided into paragraphs using Chapterize[4], and paragraphs longer than 200 words are removed. In situations where a continuous dialogue takes place over paragraph boundaries, we group them into a single paragraph. Each paragraph was annotated with one of the three aforementioned writing modes using Amazon Mechanical Turk (AMT). In addition, we add a fourth category, *Uncertain*, to encompass cases where the writing mode is unclear or does not fit well into the three main modes. All annotators were native English speakers, and three annotators were assigned to each paragraph. Paragraphs were assigned the majority label, or marked as *uncertain* in cases where each annotator provided a different label. We continued the annotation process until we had approximately 1,000 instances labeled and balanced across the three main modes (Table 1).

Writing Mode Classifier While it is possible to use the collected data to train a model, the relatively small pool of examples may cause the model to be sensitive to other text characteristics unrelated to the writing mode. To help alleviate this problem, we train a writing mode classifier and employ it to predict writing modes on a larger collection of texts. We experiment with training three separate classifiers, each trained by fine-tuning a different PLM (BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), or RoBERTa (Liu et al., 2019)) on the **FRIDE** dataset. We randomly sample 300 instances from each type of writing mode and divide them using a 1000/100/100 train/dev/test split, with an equal number of each label in each split. An evaluation of these models (Table 2) shows that all models perform similarly. The RoBERTa-based model was used as the final writing mode classifier throughout the remainder of this paper.

[2]https://www.gutenberg.org
[3]https://www.smashwords.com
[4]https://github.com/JonathanReeve/chapterize

FRIDE-XL Dataset In order to construct a larger dataset of writing modes suitable for training mode-conditional text generation models, we utilize the classifier trained in the preceding section on a larger set of texts, extending the previous text to 5,946 fiction books from Project Gutenberg. We leverage the writing mode classifier to assign a writing mode label to each paragraph of books and randomly select 362,880 paragraphs. We refer to this dataset as **FRIDE-XL**. Additional statistics of the dataset are provided in Table 7 of Appendix.

To test the accuracy of the classifier on the **FRIDE-XL** dataset, we ask three additional annotators to label 150 instances taken from the **FRIDE-XL** dataset. The results show the inter-rater agreement between annotators is 0.73. The predictions of the trained classifier agree with the majority label provided by the annotators in 85.1% instances. In terms of accuracy, classifier F1 on **FRIDE-XL** is 83.9, compared to 85.2 on the **FRIDE** dataset. It is important to reiterate that text often reflects multiple writing modes to varying degrees, and so some disagreement is inherent in the task. We find that the writing mode predicted by the classifier fails to match any label provided by human annotators in only 4.8% of cases. In Section 4.2 we show that this is indeed sufficient accuracy for producing the data and models necessary for generating mode-specific text.

3 Models

We evaluate writing mode as a control factor on three different PLM architectures: BART, (Fan et al., 2018), T5 (Raffel et al., 2020), and GPT2 (Radford et al., 2018). All models have been used previously for text generation but differ in ways that may impact their ability to adhere to the conditioning information and the quality of the generated text. For instance, the comparatively larger size and contextual window of GPT2 has made it a common choice for story generation with long text (Wang et al., 2021; Clark and Smith, 2021; Akoury et al., 2020), but smaller models like T5 show great controllability (Clive et al., 2021). We assess each of these three models, fine-tuning them to reconstruct paragraphs from the **FRIDE-XL** dataset.

For training conditional text generation models, we follow an established paradigm of summarization, conditioning, and reconstruction, as used by Sun et al. (2020). First, each paragraph is summarized using an existing summarization model.

	Model Inputs			Quality				Controllability (Accuracy)		
	L	**S**	**M**	**PPL** ↓	**B4** ↑	**RL** ↑	**BS** ↑	**Dialogue**	**Action**	**Description**
GPT2				24.41	0.95	15.02	45.61	73.33	20.28	21.67
FIST		✓		23.68	0.99	15.41	45.97	85.00	41.11	46.94
PPLM			✓	24.10	1.07	14.50	42.84	93.05	32.22	49.72
GPT2	✓			**19.29**	1.06	15.74	46.33	72.50	22.78	22.50
	✓	✓		18.84	1.14	15.97	46.70	83.33	38.61	45.56
	✓		✓	20.29	1.09	16.00	46.82	97.78	67.78	71.11
	✓	✓	✓	19.89	1.16	16.10	47.28	98.06	75.00	79.72
T5	✓			24.98	1.14	16.22	46.42	67.78	25.28	23.61
	✓	✓		23.80	1.21	16.32	46.78	80.56	47.78	46.67
	✓		✓	26.12	1.16	16.30	47.07	**99.44**	85.00	78.33
	✓	✓	✓	25.06	**1.20**	**16.52**	47.10	98.33	83.61	83.06
BART	✓			23.87	1.07	16.19	46.32	69.44	19.17	18.33
	✓	✓		23.49	1.17	16.33	47.12	86.67	47.78	48.89
	✓		✓	25.44	1.11	16.25	**47.30**	98.06	82.50	**88.06**
	✓	✓	✓	24.24	**1.20**	16.27	**47.30**	97.78	**85.56**	82.78

Table 3: Automatic evaluation on quality and controllability as model inputs (summaries (**S**), length (**L**), and writing modes (**M**)) vary. Quality is evaluated by perplexity (**PPL**), BLEU-4 **(B4)**, ROUGE-L (**RL**), BERTScore (**BS**), and controllability is measured by the accuracy of the generated stories matching the specified writing modes when the writing modes (**M**) are specified as **Dialogue**, **Action**, and **Description**. The inputs such as summaries (**S**), length (**L**), and writing modes (**M**) for the evaluation of quality and controllability are respectively inferred from the leading context and the target stories.

Here we use the narrative text summarization proposed in Kryscinski et al. (2021), and decode using beam search with a beam size of 5 as in that work. We then fine-tune a PLM to reconstruct the original paragraph, conditioning on the summary. In this way, the summary acts as a semantic control: the trained model accepts user summaries and attempts to expand upon them to generate a longer paragraph, embellishing missing and less important details in a reasonable way.

Other forms of information can also be added to the summaries to function as additional controls. The conditioning factors provided to models are:

- Summary, generated from the paragraph by a pre-trained model as shown in Appendix.
- Context, the preceding paragraph.
- Length, the number of tokens in the paragraph divided into ten equally-sized bins.
- Writing Mode, the mode assigned to the paragraph by the classifier as described in Sec. 2.

For T5 and BART, the training methodology is straightforward: we concatenate the controlling information and use it as input to the encoder, training the decoder to generate the original paragraph. For GPT2, which has only a decoder, we concatenate the conditioning information as prompts.

4 Automatic Evaluation

In this section, we study the influence of model inputs (e.g., summaries, length, and writing modes) on the quality of text, and assess to what extent the writing modes of text can be controlled, as measured by automatic evaluation metrics.

4.1 Baseline Models

In addition to ablations of our proposed models, we compare against three baseline systems:

GPT2 We finetune GPT2 (Radford et al., 2018) identically to our proposed system, but using only the preceding paragraph and without other inputs.

PPLM Dathathri et al. (2020) employ an attribute classifier to guide the pretrained language model to generate text with specified attributes. To adapt the PPLM to our task, we train a writing mode classifier as the attribute classifier on the **FRIDE** dataset. As the writing modes of preceding paragraphs would interfere with the classifier, the PPLM does not take the preceding paragraphs as context.

FIST Fang et al. (2021) propose a system which utilizes keywords instead of summaries to sketch the semantic content of the desired stories. As there is no prompt in our dataset, following their idea, we infer the keywords from the leading context (the

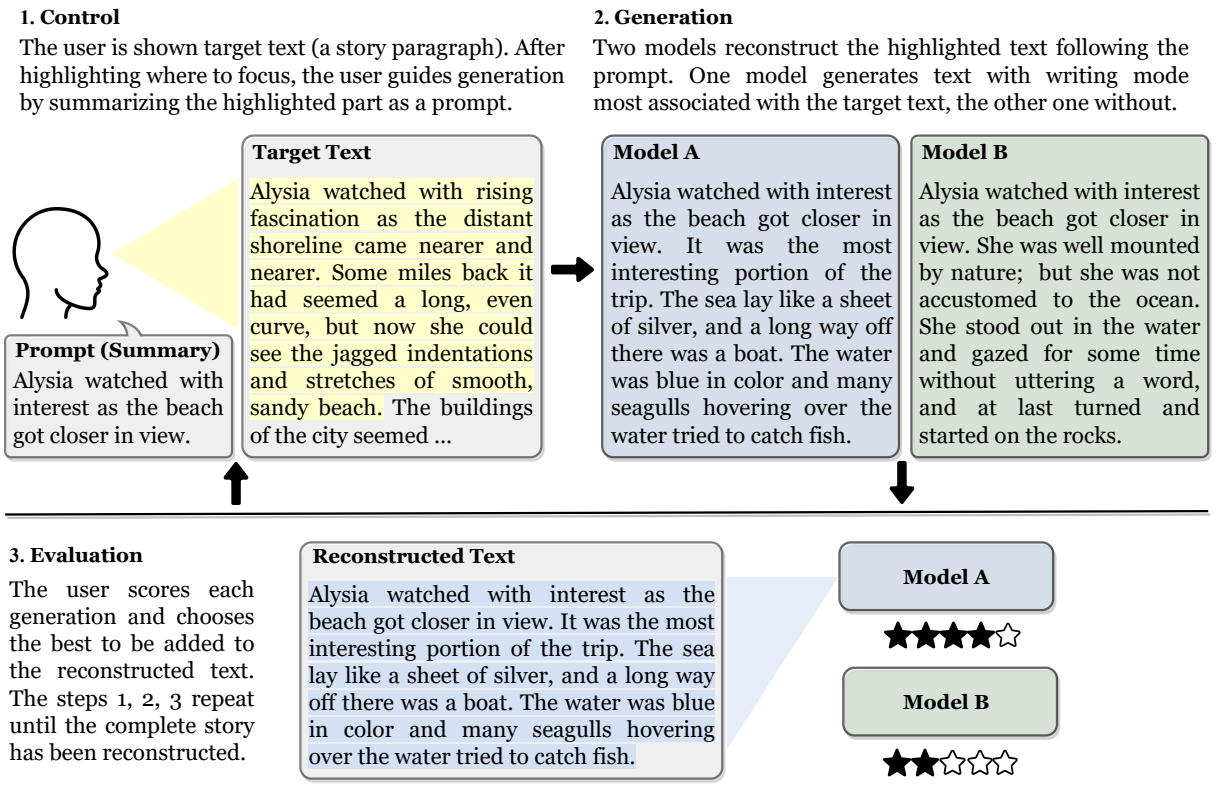

Figure 2: A turn-based text reconstruction game. Users aim to reconstruct the target text solely through interaction with story generation models, providing prompts (summaries) to guide generation, and choose between a number of options to continue the text. Each option is generated by a different model. Users report satisfaction at each turn, and the reconstructed text is used in further evaluation.

preceding paragraphs) and then generate stories conditioning on the context and keywords.

4.2 Results

We evaluate the models along three axes: fluency, similarity, and controllability, using the test set of the **FRIDE-XL** dataset. The results of our automatic evaluation are shown in Table 3.

Fluency We evaluate fluency using perplexity computed by the pre-trained GPT2 model. We find that there is an average 0.8 decrease in perplexity when summaries are added and 1.2 increase when writing modes are added.

Similarity Our task is very similar to summarization, so we adopt the same evaluation metrics including BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and BERTScore (Zhang et al., 2020). It is also quite common to include these automatic measures in narrative text generation work, despite their known flaws. We observe a consistent improvement across all models as the amount of conditioning context increases, and the models using writing mode factors outperform those without.

Controllability Lastly, we evaluate the controllability of mode-controlled models. For each paragraph, a target writing mode is chosen using the writing mode classifier, and used as conditioning for a text generation model. The classifier is then used to predict the writing mode of the generated text, and we measure the accuracy of generating stories with the specified writing modes.

On average, including writing mode as condition improves the accuracy of generating text which is classified as that mode, but the effect varies drastically by the specific mode. For Action and Description modes, the inclusion of writing mode conditioning improves accuracy on average by 45.4% and 45.6%, respectively, compared to their non-mode counterparts. For dialogue, the improvement is 21.5%, relatively lower.

It is interesting to note that the inclusion of summaries to the length-only model results in significant improvements to the controllability of the text. This implies that the pre-trained models are able to naturally infer the intended writing mode from the summaries to some degree, with modest accuracy ($\sim 44\%$) on average for Action and De-

Model	Writing Mode Control	Selection Percentage			Satisfaction Score		
		Dialogue	Action	Description	Dialogue	Action	Description
GPT2	✗	31	44	32	2.27	2.84	2.67
	✓	**69**	**56**	**68**	**2.96**	**3.27**	**3.59**
BART	✗	36	39	43	2.82	2.64	3.10
	✓	**64**	**61**	**57**	**3.28**	**3.11**	**3.34**
T5	✗	47	36	44	3.19	2.91	3.22
	✓	**53**	**64**	**56**	**3.24**	**3.36**	**3.55**

Table 4: Human evaluation by authors, showing **Selection Percentage** and **Satisfaction Score** between generated text w/ and w/o writing mode conditioning, during the text reconstruction game. **Dialogue**, **Action**, and **Description** are the writing modes of the target text.

scription modes, and up to 86% on Dialogue with BART. Summaries may contain some cues about the intended modes, especially, the summaries for Dialogue have strong cues (*said*, *replied*, *argued*, ...) in most cases. However, the consistent large margins of accuracy scores when conditioning on writing modes illustrates the effectiveness of modes as a control factor.

5 Human Evaluation

In the preceding section, we used automatic measures to show consistent improvements in similarity and controllability when conditioning text generation models on writing modes. In this section, we assess whether these improvements translate into higher author satisfaction and story quality as judged by human participants.

Text Reconstruction Game To evaluate writing mode-controlled models within human-machine collaborative writing, we devise a story generation game (Figure 2). Using a web interface, participants are presented with a target story, drawn from **FRIDE-XL**, and asked to reconstruct it through interaction with the trained models. Each session utilizes two models, each using the same PLM, one trained to condition on writing modes, and one that does not.

At each round, (1) the author highlights a selection of text from the target story, and provides a text summary which serves as a rough sketch of the story described therein. Constraints on the summary length and a time limit on each session force the author to extract essential elements of the text and prevent them from simply specifying the entire text as the summary. (2) The models then each return a sample of generated text, and (3) the author is asked to score each in terms of overall satisfaction (ordinal, 5-point scale), and choose one to continue the story. The source of the suggestions (model name) is not given to the participant, and they are presented in random order. The above process repeats until the author has attempted to recreate the entire target text. The author is then asked to provide feedback on overall satisfaction with the story and interface, the efficiency of using the interface over generating the text from scratch, and what errors were present in the generated text.

Our approach takes inspiration from previous human-in-the-loop evaluations (Akoury et al., 2020; Clark and Smith, 2021) where authors are asked to construct stories by collaborating with story generation models, but do so freely, without being provided any specific direction for how the story should unfold. Therefore we cannot observe the author's intention, and it is difficult to ascertain how closely the generated stories match the author's original intention, or the extent to which bias may be affecting the model scores when satisfaction with the models is self-reported.

Our solution to these issues is to present the author with a target text they must attempt to reconstruct. The target text serves as a substitute for the author's intention, allowing us to more objectively measure the differences between the generated story and the target. While this design shifts the nature of interaction away from a more creative use case, we argue that increased awareness of the models' limitations provides a worthwhile trade-off when used strictly as a means of evaluation.

Using the unique properties of our proposed evaluation design, we evaluate: (1) the effectiveness of writing mode control when generating text as an author, (2) their effectiveness from the perspective of a reader, comparing generated stories with target stories in a blind study, and (3) the types of errors made where authors were unable to exert desired

	Collab.	Auto.	P	+WM	-WM	P
Plot Similarity						
GPT2	**3.60**	3.04	0.01	**3.42**	3.30	0.51
BART	3.59	**3.66**	0.71	**3.56.**	3.22	0.06
T5	**3.49.**	3.05	0.01	**3.48.**	3.10	0.02
Style Similarity						
GPT2	**3.21.**	3.05	0.43	**3.48**	3.15	0.02
BART	**3.80.**	3.59	0.28	**3.71**	3.30	0.02
T5	**3.56**	3.19	0.03	**3.48**	3.08	0.02

Table 5: The similarity scores (5 points scale) in plots and style between the reconstructed text and the target ones. **Collab.** and **Auto.** refer to the text collaboratively reconstructed with the users and automatically generated by the models. **+WM** and **-WM** are the text automatically reconstructed by models with or without writing modes. **P** is the P-Value of significance test.

control over the models.

5.1 Evaluation by Authors

We randomly sample 36 target stories (each annotated with a writing mode) from **FRIDE-XL**, such that there is an equal number of stories from each mode and each genre. For the writing mode-controlled model, we use the mode annotated in the dataset as the true mode (users are not asked to specify a mode explicitly). The authors are recruited via crowdsourcing on AMT, and each target story is reconstructed using the proposed interface, with each story given to three annotators.

Do authors prefer the suggestions from writing mode-controlled models? We measure authors' preferences for models by the win/loss rate for which model was selected to continue the story at each step, and the average satisfaction score of each model's suggestions. The results are presented in Table 4. Suggestions of writing mode models are preferred consistently, irrespective of model type, and are chosen on average 22.2% more than suggestions from uncontrolled models. In terms of satisfaction score, using writing mode as a conditioning factor improves satisfaction by 0.5 / 5.0 points on average, an improvement consistent across all writing modes. We conclude that the strong and consistent improvements when using writing mode-conditional models demonstrate they are an effective control for story generation.

5.2 Evaluation by Readers

Evaluation by authors using the story reconstruction game interface provides compelling evidence for the effectiveness of writing modes as a control, but how close to the author's intention are the stories generated using the writing interface? By providing authors with a target story, new evaluation methods are possible, such as having the stories scored by a separate group of participants, in order to avoid any bias from self-reporting.

Here we evaluate the generated texts via similarity with the target texts, in two different scenarios. We enlist 197 human participants from AMT, none of which participated in the story generation task, to serve as readers. We present each reader with three stories: the target story, the generated story, and a baseline story. Readers are asked to score the similarity between the target text and the other stories on an ordinal 5-point scale, in terms of plot and style.

How similar to the target text are the generated stories? The generated stories are written interactively via the writing interface, where authors have access to suggestions from mode and non-mode models. The baseline stories are generated in a purely automated manner using the model to predict the entire paragraph from a summary (generated by the same summarization model we use in the fine-tuning process), together with the writing modes provided from the dataset annotations.

The left side of Table 5 shows the results. We observe higher similarity scores between the target text and the stories generated via collaborative writing in almost all cases, an average increase of 0.31 on plot, and 0.25 on style. On one hand, this is an expected conclusion given that the collaborative process allows authors to select the best of two automatic suggestions at each point, and one may expect the quality of collaborative writing stories to be strictly better than automatic ones. However, we do note trends based on model type; BART in particular scores high (3.80) in terms of similarity on style.

How effective is writing mode control for automatic generation While our focus is primarily on the use of additional controls in collaborative story writing, we also measure the effectiveness of writing modes as a useful control in a purely automatic sense. Using the story generation models, as above, we generate full stories without writing mode conditioning and contrast them with the previously generated stories generated automatically using writing modes. We report the similarity of

Error Type	GPT2	BART	T5
Missing Information	61	65	32
Irrelevant Information	45	40	22
Wrong Information	8	8	2
Incoherence	22	22	22
Inconsistency	22	13	20
Disfluency	15	27	18
Repetition	0	0	0

Table 6: The counts of each error occurring in the reconstructed text across different models.

each to the target text in the right side of Table 5.

We again find that writing mode-controlled models significantly improve the control of the generated text, producing stories closer to the target text in both style and content. The addition of a writing mode control improves average similarity to target text by 0.28 on plot and 0.38 in style. Although writing modes are more inherently tied to the style of the text, it is interesting to observe improvements in plot similarity as well. This may indicate that the mode has a positive effect in pressuring the model to focus on summary plot points, via explicitly disentangling these factors. However, we leave further analysis of this phenomenon to future work.

5.3 Error Analysis

To understand what errors are likely to occur in the generated stories we ask annotators to identify aspects of the generated text which differed from the target, from a set of pre-determined categories (Table 6). The most frequent errors are the *missing information* and *irrelevant information*, suggesting that the models extrapolated from the summaries in undesirable ways.

Note that the counts of most errors made by T5 are appreciably lower than those made by GPT-2 or BART, yet BART has higher similarity scores in both plot and style. We infer that some errors made by T5 must play a more important role in overall similarity, and that missing or irrelevant information must not play a crucial role in similarity metrics.

6 Related Work

Our work is based on prior research in computational modeling for story generation. Early approaches to automatic story generation relied on graph-based planning and hand-crafted rules to structure narratives (Meehan, 1977; Callaway and Lester, 2002; Riedl and Young, 2004; Li et al., 2013). More recent works generate stories by finetuning on large-scale PLMs (See et al., 2019) to improve its fluency and incorporating structured knowledge such as planned events (Chen et al., 2021; Fang et al., 2021; Li et al., 2022), summaries (Yao et al., 2019; Tan et al., 2021; Sun et al., 2020), and external knowledge (Guan et al., 2019; Xu et al., 2020b; Guan et al., 2020) to enhance its coherence and consistency. Our story generation models are also finetuned on the large-scale PLMs to generate text following the given summaries.

Our work bears similarity to work on controllable story generation, which aims to control different attributes of stories such as the sentiments (Luo et al., 2019; Kong et al., 2021b), genres (Cho et al., 2022), intention (Sun et al., 2021), and characters (Lee and Jung, 2020; Xu et al., 2020a; Liu et al., 2020b) of stories. However, these attributes are largely unchanging throughout the story, while we focus on writing mode, a more dynamic attribute of text. Thus our work is also more inherently related to interactive story generation where the author works closely with the model to craft text on a comparatively finer level of granularity (sentences or paragraphs).

Finally, our evaluation method is inspired by work on human-in-the-loop storytelling (Roemmele and Gordon, 2015; Samuel et al., 2016; Clark et al., 2018; Goldfarb-Tarrant et al., 2019; Brahman et al., 2020), where the authors are asked to work in concert with story generation models, curating their suggestions to craft the final story. Human-in-the-loop evaluations overcome many of the shortcomings of automatic evaluations, which capture rough statistics, but may be unaware of important errors in plot development and story continuity (Sagarkar et al., 2018). By asking users to select between models' suggestions, we can instead gain a more accurate picture of which system improvements yield real benefits to a potential human-machine writing collaboration (Akoury et al., 2020; Clark and Smith, 2021; Khashabi et al., 2021). Our approach is similar to this, but the addition of a target text allows us to examine the difference between the generated and intended text, which we argue is a more important comparison when dealing specifically with understanding *control*.

7 Conclusion

In this work, we introduced writing modes as a control for human-machine collaborative writing

scenarios and showed that training models to condition on writing modes resulted in stories that were closer to targets. Both the automatic and human evaluation shows that the writing modes of text are effectively controlled, authors prefer text suggestions made by writing mode-controlled models, and readers score stories to be more similar to targets in terms of both plot and style. To verify the hypothesis, we collected **FRIDE** and **FRIDE-XL**, datasets of narrative text annotated with writing modes, which we released to help facilitate further research in writing modes and fine-grained control for storytelling. In future work, we wish to apply reconstruction-based evaluation for other factors of human-machine storywriting, and incorporate a dynamic use of writing modes into fully automatic hierarchical story generation models.

Limitations

This work is subject to known biases in the dataset used throughout this work. Due to existing copyrights on most contemporary examples of professional narrative fiction, researchers often turn to works in the public domain, as we do here. While many public domain novels are literary classics, the lack of comparable contemporary work results in models which are biased towards reproducing older works, both in terms of style and content. More contemporary approaches to writing style are not represented in our work, and plot points may be biased by the worldview of the authors at the time.

Acknowledgements

This paper is based on results obtained from a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This work is also partially supported by JSPS KAKENHI Grant Number JP19H05692. We sincerely thank Yusuke Mori, Yang Zhao, and the anonymous reviewers for their advice and feedback on earlier drafts of this work.

References

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A dataset and evaluation platform for machine-in-the-loop story generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6470–6484. Association for Computational Linguistics.

Faeze Brahman, Alexandru Petrusca, and Snigdha Chaturvedi. 2020. Cue me in: Content-inducing approaches to interactive story generation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, pages 588–597. Association for Computational Linguistics.

Charles B Callaway and James C Lester. 2002. Narrative prose generation. *Artificial Intelligence*, 139(2):213–252.

Hong Chen, Raphael Shu, Hiroya Takamura, and Hideki Nakayama. 2021. Graphplan: Story generation by planning with event graph. In *Proceedings of the 14th International Conference on Natural Language Generation, INLG 2021, Aberdeen, Scotland, UK, 20-24 September, 2021*, pages 377–386. Association for Computational Linguistics.

Jin-Uk Cho, Min-Su Jeong, JinYeong Bak, and Yun-Gyung Cheong. 2022. Genre-controllable story generation via supervised contrastive learning. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2839–2849. ACM.

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative writing with a machine in the loop: Case studies on slogans and stories. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces, IUI 2018, Tokyo, Japan, March 07-11, 2018*, pages 329–340. ACM.

Elizabeth Clark and Noah A. Smith. 2021. Choose your own adventure: Paired suggestions in collaborative writing for evaluating story generation models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3566–3575. Association for Computational Linguistics.

Jordan Clive, Kris Cao, and Marek Rei. 2021. Control prefixes for text generation. *CoRR*, abs/2110.08329.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Angela Fan, Mike Lewis, and Yann N. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 889–898. Association for Computational Linguistics.

Le Fang, Tao Zeng, Chaochun Liu, Liefeng Bo, Wen Dong, and Changyou Chen. 2021. Outline to story: Fine-grained controllable story generation from cascaded events. *CoRR*, abs/2101.00822.

Seraphina Goldfarb-Tarrant, Haining Feng, and Nanyun Peng. 2019. Plan, write, and revise: an interactive system for open-domain story generation. 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, Demonstrations*, pages 89–97. Association for Computational Linguistics.

Jian Guan, Fei Huang, Minlie Huang, Zhihao Zhao, and Xiaoyan Zhu. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Trans. Assoc. Comput. Linguistics*, 8:93–108.

Jian Guan, Xiaoxi Mao, Changjie Fan, Zitao Liu, Wenbiao Ding, and Minlie Huang. 2021. Long text generation by modeling sentence-level and discourse-level coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6379–6393. Association for Computational Linguistics.

Jian Guan, Yansen Wang, and Minlie Huang. 2019. Story ending generation with incremental encoding and commonsense knowledge. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6473–6480. AAAI Press.

Stephen Halliwell and Aristotle. 1998. *Aristotle's Poetics*. University of Chicago Press, Chicago.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel S. Weld. 2021. GENIE: A leaderboard for human-in-the-loop evaluation of text generation. *CoRR*, abs/2101.06561.

M. Klaassen. 2015. *Fiction-Writing Modes: Eleven Essential Tools for Bringing Your Story to Life*. Bookbaby.

Xiangzhe Kong, Jialiang Huang, Ziquan Tung, Jian Guan, and Minlie Huang. 2021a. Stylized story generation with style-guided planning. *arXiv preprint arXiv:2105.08625*.

Xiangzhe Kong, Jialiang Huang, Ziquan Tung, Jian Guan, and Minlie Huang. 2021b. Stylized story generation with style-guided planning. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2430–2436. Association for Computational Linguistics.

Wojciech Kryscinski, Nazneen Fatema Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir R. Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization. *CoRR*, abs/2105.08209.

O-Joun Lee and Jason J. Jung. 2020. Story embedding: Learning distributed representations of stories based on character networks (extended abstract). In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 5070–5074. ijcai.org.

Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. 2013. Story generation with crowdsourced plot graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27.

Qintong Li, Piji Li, Wei Bi, Zhaochun Ren, Yuxuan Lai, and Lingpeng Kong. 2022. Event transition planning for open-ended text generation. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3412–3426. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Danyang Liu, Juntao Li, Meng-Hsuan Yu, Ziming Huang, Gongshen Liu, Dongyan Zhao, and Rui Yan. 2020a. A character-centric neural model for automated story generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1725–1732.

Danyang Liu, Juntao Li, Meng-Hsuan Yu, Ziming Huang, Gongshen Liu, Dongyan Zhao, and Rui Yan. 2020b. A character-centric neural model for automated story generation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI*

Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 1725–1732. AAAI Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Fuli Luo, Damai Dai, Pengcheng Yang, Tianyu Liu, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. Learning to control the fine-grained sentiment for story ending generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6020–6026. Association for Computational Linguistics.

James R Meehan. 1977. Tale-spin, an interactive program that writes stories. In *Ijcai*, volume 77, page 9198.

Jessica Morrell. 2006. *Between the lines: Master the subtle elements of fiction writing*. Penguin.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

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.

Mark Owen Riedl and R Michael Young. 2004. An intent-driven planner for multi-agent story generation. In *Autonomous Agents and Multiagent Systems, International Joint Conference on*, volume 2, pages 186–193. IEEE Computer Society.

Melissa Roemmele and Andrew S. Gordon. 2015. Creative help: A story writing assistant. In *Interactive Storytelling - 8th International Conference on Interactive Digital Storytelling, ICIDS 2015, Copenhagen, Denmark, November 30 - December 4, 2015, Proceedings*, volume 9445 of *Lecture Notes in Computer Science*, pages 81–92. Springer.

Manasvi Sagarkar, John Wieting, Lifu Tu, and Kevin Gimpel. 2018. Quality signals in generated stories. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 192–202. Association for Computational Linguistics.

Ben Samuel, Michael Mateas, and Noah Wardrip-Fruin. 2016. The design of writing buddy: A mixed-initiative approach towards computational story collaboration. In *Interactive Storytelling - 9th International Conference on Interactive Digital Storytelling, ICIDS 2016, Los Angeles, CA, USA, November 15-18, 2016, Proceedings*, volume 10045 of *Lecture Notes in Computer Science*, pages 388–396.

Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019. Do massively pretrained language models make better storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pages 843–861. Association for Computational Linguistics.

Simeng Sun, Wenlong Zhao, Varun Manjunatha, Rajiv Jain, Vlad I. Morariu, Franck Dernoncourt, Balaji Vasan Srinivasan, and Mohit Iyyer. 2021. IGA: an intent-guided authoring assistant. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5972–5985. Association for Computational Linguistics.

Xiaofei Sun, Chun Fan, Zijun Sun, Yuxian Meng, Fei Wu, and Jiwei Li. 2020. Summarize, outline, and elaborate: Long-text generation via hierarchical supervision from extractive summaries. *CoRR*, abs/2010.07074.

Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric P. Xing, and Zhiting Hu. 2021. Progressive generation of long text with pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4313–4324. Association for Computational Linguistics.

Wei Wang, Piji Li, and Hai-Tao Zheng. 2021. Consistency and coherency enhanced story generation. In *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I*, volume 12656 of *Lecture Notes in Computer Science*, pages 694–709. Springer.

Feifei Xu, Xinpeng Wang, Yunpu Ma, Volker Tresp, Yuyi Wang, Shanlin Zhou, and Haizhou Du. 2020a. Controllable multi-character psychology-oriented story generation. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1675–1684. ACM.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020b. MEGATRON-CNTRL: controllable story generation with external knowledge using large-scale language models. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 2831–2845. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7378–7385. AAAI Press.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

8 Appendix

8.1 Writing Mode Classifier

To study how many samples are enough to train the classifiers, we test the classifiers trained on the subsets of the FRIDE dataset. The subset size increases from 0 to 1000. The results in Figure 3 show the classifiers reach the best performance when the subset size is around 200.

When training the classifiers, we use the optuna [5] to do a hyper-parameter search. We run hyper parameter search 20 times and try to search the optimal value for learning rate (in range 1e-5∼5e-5), batch size (in range 2∼8), and training epochs (in range 2∼10). The optimal value for learning rate is 1e-5, for epoch is 10, and for batch size is 5. The reported test results are the average of 10 trails with different random seeds.

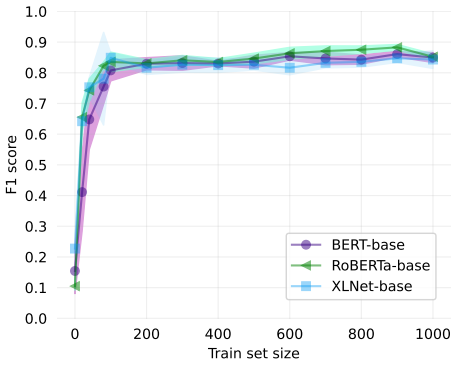

Figure 3: The F1 score of 3 classifiers trained on sub-train sets with different size.

8.2 Writing Mode Distribution

The writing mode is a kind of style to tell stories, and thus, it could be related to the genre of fictions. To verify that the writing modes are a style independent of the genres, we select 9 frequent genres and analyze the distribution of the writing modes within each genre. The results in Figure 4 show that the writing modes have similar distribution across different genres, demonstrating that the distribution of writing modes is irrelevant to the category of genre. Thus, the writing mode is a style independent of the genre.

8.3 Training Settings

The story generation models are finetuned from three types of pretrained language models such as BART, GPT2, and T5. We utilize the large version

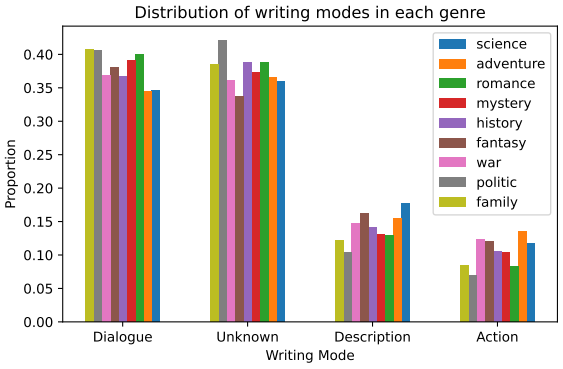

Figure 4: The writing mode distribution in each genre.

of the pretrained models such as “bart-large",“gpt2-large", and “t5-large", for better performance, and these models are downloaded through huggingface [6] model library. The story generation models are trained on 8 Tesla A100 with a learning rate 4e-5 and a batch size 32 for 5 epochs about 24 hours. The warm up steps for all models are the first 10% steps of the total steps. During inference phase, we adopt nuclear sampling method with top_p being 0.9 recommended by Holtzman et al. (2020).

	Train	Valid	Test
#Paragraph	360K	1.44K	1.44K
#Dialogue	131.196K	0.493K	0.36K
#Description	54.005K	0.259K	0.36K
#Action	30.478K	0.103K	0.36K
#Uncertain	144.321K	0.585K	0.36K
#Token of Paragraph	111.3	111.4	111.0

Table 7: The statistics of FRIDE-XL Dataset.

	ROUGE-1	ROUGE-2	ROUGE-L
Paper	22.2	4.8	16.9
Our	21.5	4.4	16.5

Table 8: The evaluation for the performance of the summarization model.

8.4 Summarization Model

Most prior works focus on the summarization of news articles, which is a domain different from the narrative text in books. Recently, Kryscinski et al. 2021 collect a summarization dataset for the narrative text in books. Each book in the dataset is summarized in different levels including paragraph-level, chapter-level, and book-level. We run their

[5] https://optuna.readthedocs.io/en/stable/

[6] https://huggingface.co

codes[7] and train a paragraph-level summarization model. The quality of summaries is evaluated by ROUGE scores as shown in the Table 8. The Table shows that we basically reproduce their results in the paper. However, better summarizers for narrative text could greatly improve the output of our story generation models. In particular, while the summaries are often correct, the focus of what aspects of the narrative text is summarized is not always the most appropriate, and improving this is a clear direction for future work. As for the summary size, the trained model summaries the original paragraph ($\sim$116 tokens) into sentences ($\sim$26 tokens). The average compression rate is 18.2%.

[7] https://github.com/salesforce/booksum