Advancing Precise Outline-Conditioned Text Generation with Task Duality and Explicit Outline Control

Anonymous ACL submission

Abstract

Existing works on outline-conditioned text generation typically aim to generate text using provided outlines as rough sketches, such as keywords and phrases. However, these approaches make it challenging to control the quality of text generation and assess consistency between outlines and generated texts due to lack of clarity and rationality of the rough outlines. In this paper, we introduce a novel text generation task called ***Precise Outline-conditioned Generation***, which requires generating stories based on *specific*, *sentence-level* outlines. To facilitate research on this task, we construct two new datasets, **WPOG** and **CDM**. We provide strong baselines based on fine-tuning models such as BART and GPT-2, and evaluating zero-shot performance of models such as ChatGPT and Vicuna. Furthermore, we identify an issue of **imbalanced utilization of the outline information** in the precise outline-conditioned generation, which is ubiquitously observed across fine-tuned models and zero-shot inference models. To address this issue, we propose an **explicit outline utilization control approach** and a novel framework that **leverages the task duality between summarization and generation**. Experimental results show that the proposed approaches effectively alleviate the issue of imbalanced outline utilization and enhance the quality of precise outline-conditioned text generation for both fine-tuning and zero-shot settings.[1]

1 Introduction

Outline-conditioned text generation is a challenging but important task, which provides

---

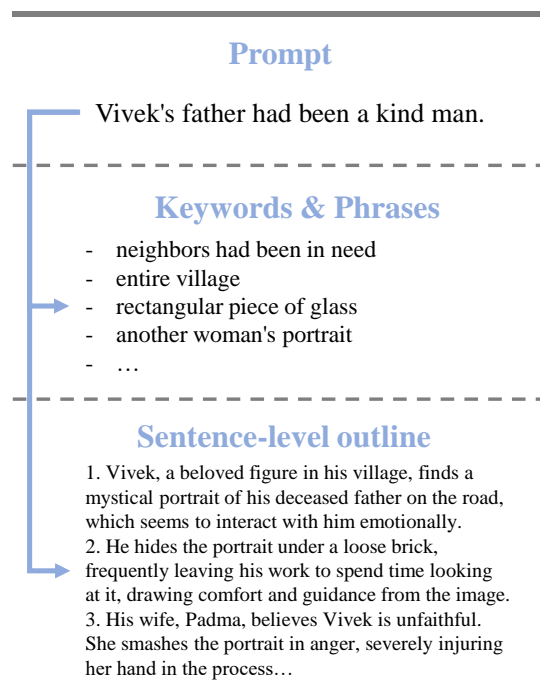[1]We will release the source code and datasets to facilitate future research.

Figure 1: The comparison between different forms of outlines. Given the prompt (e.g., title), the outline could be formulated as a) **rough** outlines: a list of keywords or phrases, and b) **precise** outlines: salient sentence-level statements.

structured conditions (outlines) that aim to enhance content quality, customization, and generation efficiency for a variety of downstream text generation applications, e.g., story-telling and long-form question-answering. It requires the model to not only generate fluent and coherent text but also ensure that the generated text aligns with the content, structure, and properties (e.g., style) specified in the outline. Existing works (Fan et al., 2018; Rashkin et al., 2020; Fang et al., 2021) mainly regard a set of keywords or phrases as the "rough outline". The rough outline is usually extracted via heuristic methods (Rose et al., 2010; Campos et al., 2018) and is overly abstractive, making it difficult to guarantee either quality or rationality

of the rough outline, which in turn hampers quality of the generated text conditioned on the outline. Moreover, the ambiguity and poor readability of the rough outline also complicate the evaluation of consistency between the outlines and generated text (e.g., simply asking the evaluator "Which text is better at utilizing the keywords?" as in (Rashkin et al., 2020) could be quite challenging to answer).

To address these limitations of text generation conditioned on rough outlines, we introduce a novel task called ***Precise Outline-conditioned Generation***. As shown in Figure 1, this task takes the specific, sentence-level outline rather than keywords and phrases, which could improve the readability, clarity, and rationality of outlines and hence the quality of generated text and could better facilitate assessing consistency between the outline and generated text. Compared to text generation conditioned on rough outlines, our proposed task poses greater challenges in three aspects: a) **Controllability:** Precise outlines inherently encode properties (e.g., style and attributes) that vary drastically with scenarios (e.g., stories, news reports). Models need to grasp the properties and treat them as pre-given constraints to control text generation. b) **Faithfulness:** The generated text needs to be faithful to the outline in terms of not only content but also relations among content in the outline. c) **Structure:** The generated text needs to maintain a reasonable structure so that the key points in the generated text are well organized.

To facilitate research on this new task, we provide two datasets **WPOG** and **CDM** from storytelling and news report scenarios, respectively. We study how competitive models address these challenges by fine-tuning widely used pre-trained language models (LMs) such as BART (Lewis et al., 2020) and GPT-2 (Radford et al., 2019) for inference and also zero-shot inference with top-performing large language models (LLMs) such as ChatGPT (Ouyang et al., 2022) and Vicuna (Chiang et al., 2023)[2]. The two datasets exhibit two drastically different styles, so we could evaluate how well

models understand the inherent style and control generation conditioned on the style, i.e., handling the **controllability** challenge. We observe a common issue of imbalanced utilization of the outline across fine-tuning and zero-shot settings, which affects the structure of the generated text. Specifically, the outline information is mostly used at the beginning of text generation and does not effectively guide the rest of the generation process. This issue indicates that the models struggle with the challenges of maintaining **faithfulness** and **structure**. We hypothesize that the imbalanced outline utilization issue may be caused by a combination of exposure bias due to teacher forcing (the model relies too heavily on previously predicted words) and decoding strategies (the model assigns the highest probabilities to the correct outline information), leading to **early depletion** of the outline. In order to accurately assess the faithfulness and structure quality of the generated text, we propose three metrics: **distribution variation**, **peak-value distance**, and **consistency degree**.

We propose two general approaches to mitigate this imbalance issue. The first method explicitly controls the usage of the given outline throughout the generation process. The second method exploits the task duality between precise outline-conditioned generation and summarization and implements a dual learning strategy. Experimental results on the two datasets with automatic and human evaluations show that both approaches alleviate the imbalance issue and improve generation quality.

Our contributions can be summarized as follows: (1) We introduce a novel **Precise Outline-conditioned Generation** task. (2) We provide two datasets **WPOG** and **CDM** to support research on this task. (3) We identify the **imbalanced outline utilization issue** and propose three evaluation metrics to gauge outline utilization. (4) To alleviate this imbalanced outline utilization issue, we propose two *general* approaches, **explicit outline utilization control** and **unified dual-task learning**. (5) Experimental results demonstrate that our approaches effectively mitigate the imbalanced outline utilization issue and improve the generation quality **for both fine-tuning and zero-shot settings**.

---

[2]For LLMs with strong zero-shot inference ability (Kojima et al., 2023), in this work, we only focus on in-context learning without further instruction tuning or other fine-tuning.

2 Related Work

Controllable Text Generation Controllable text generation is a core issue in text generation, which is also a requirement in virtually every context involving text generation (Zhang et al., 2022). Previous works proposed a lot of related tasks to evaluate the controllability from coarse to fine-grained. Attribute-based generation requires the attributes-specific sentence with given topics (Dathathri et al., 2019; Wang et al., 2019), keywords (Zhang et al., 2020; Carlsson et al., 2022) or sentiments (He et al., 2020; Zhang and Song, 2022). However, the sentence-level target has less requirement for coherence or structure. For longer text generation, storytelling, and data-to-text aim to generate paragraphs or passages based on given topics and endings (Fan et al., 2018; Goldfarb-Tarrant et al., 2020; Tambwekar et al., 2018), or tables and graph (Su et al., 2021; Ribeiro et al., 2020). The closest task to ours is planning-based generation, which we call rough outline-conditioned generation. The main difference is that they primarily use keywords or phrases as conditions, which limits the requirements for generating results and assessing accuracy due to the vagueness of the conditions themselves. On the other hand, the rough outline is usually extracted by heuristic methods (e.g., RAKE (Rose et al., 2010) and YAKE (Campos et al., 2018)) to select keywords from ground truth text by word frequency and graph metrics, which also limits the quality of data.

Outline-conditioned Generation There are a variety of works that take outline-conditioned generation as the critical step to long text generation. Xu et al. (2018); Yao et al. (2019); Shao et al. (2019) adopt the Bi-LSTM as the backbone with reinforcement learning and variational inference techniques to optimize with keywords/phrases skeleton, respectively. Fan et al. (2019) modeling the semantic role labels (SRL) as outlines by the self-attention mechanism and pointer mechanism (Vinyals et al., 2015). Tan et al. (2020); Chen et al. (2021); Wang et al. (2022) use pre-trained language models to tackle with outlines in the form of keywords, event-graph, and latent variables, respectively. Sun et al. (2022) propose a method to learn outline generation by reconstructing the summary, followed by generating segment-level text and concatenating them to obtain the full text. For zero-shot inference with LLMs, Yang et al. (2022) applies the recursive prompting and revision to enhance GPT-3's zero-shot performance on long-form storytelling. Yang et al. (2023) make a detailed outliner and a controller based on FUDGE (Yang and Klein, 2021) to improve long-range plot coherence. In this paper, we consider both fine-tuning and zero-shot settings and propose unified strategies that could work for both scenarios.

3 Methods

3.1 Precise Outline-conditioned Generation

Baseline Approach Our proposed precise outline-conditioned generation task takes specific sentence-level outlines and prompts as input, and requires generating long texts that are fluent, coherent, and match the input information. Specifically, the writing prompt (e.g., title or opening) is denoted by $\mathbf{x} = \{x_1, x_2, ..., x_l\}$, where x_i denotes the i-th token in the writing prompt. The outline $\mathbf{o}$ is a set of sentences $\mathbf{o} = \{o_1, o_2, ..., o_m\}$, where o_j is the j-th sentence and o^t is the t-th token in $\mathbf{o}$. Given $(\mathbf{x}, \mathbf{o})$ as input, a model is expected to output $\mathbf{y} = \{y_1, y_2, ..., y_n\}$, where y_k denotes the k-th sentence and y^t is the t-th token in $\mathbf{y}$.

We study two distinct approaches for implementing $(\mathbf{x}, \mathbf{o}) \rightarrow \mathbf{y}$: **fine-tuning pre-trained LMs** and **zero-shot inference by LLMs**. For fine-tuning pre-trained encoder-decoder and decoder-only pre-trained LMs and then inference, the output is generated by the model θ as follows:

$$P(\mathbf{y} \mid \mathbf{o}, \mathbf{x}; \theta) = \prod_{t=1}^{n} P\left(y^t \mid y^{<t}, \mathbf{o}, \mathbf{x}; \theta\right) \quad (1)$$

An encoder-decoder model (e.g., BART) encodes the outline $\mathbf{o}$ by its encoder and treats the prompt $\mathbf{x}$ as the prefix for its decoder; whereas, a decoder-only model (e.g., GPT-2) treats both the outline and the prompt as the prefix for its decoder.

For LLM zero-shot inference, we adopt *prompt learning* to infer with the concatenation of the outline and the prompt. Details of prompt designing are in Appendix A.1.

3

Imbalanced outline utilization We identify a common issue of imbalanced outline utilization in fine-tuning both encoder-decoder and decoder-only models as well as zero-shot inference. We conduct case studies in Table 5 (Appendix E) and a similarity visualization in Figure 5 (Appendix C) to further illustrate the problem. Specifically, the models tend to repeat all the outlines at the head of the output and then continue writing while only treating the output prefix as context. This issue indicates weaknesses of the models for handling **Faithfulness** and **Structure** challenges (Section 1). Lack of constant guidance from the outline to regulate the remaining writing may cause deviations from the outline and unfaithful content, resulting in poor faithfulness. Also, with the skewed distribution of repeating all the outlines at its head, the generated text could be ill-structured. We propose three metrics to evaluate outline utilization in Section 5.1. To address this issue, we propose an explicit outline utilization control method and a unified dual learning strategy.

3.2 Explicit Outline Utilization Control

As Figure 2 illustrates, we propose an approach that explicitly controls the utilization of the outline during generation by segmenting the outline into points (taking one outline sentence as one point in this work), generating text corresponding to each point, and aggregating and refining the outputs to obtain the final result. We denote this new method as **explicit outline utilization control (OC)**. OC can be used in both fine-tuning and zero-shot inference settings.
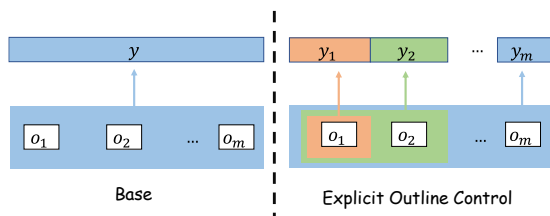

Figure 2: The proposed **Explicit Outline Utilization Control method** versus the baseline approach. We omit the prompt **x** for simplicity.

Fine-tuning We apply OC during both fine-tuning and inference. For each story **y** and its corresponding outline $\mathbf{o} = \{o_1, o_2, ..., o_m\}$ in the training set, we partition **y** into m snippets $\mathbf{s} = \{s_1, s_2, ..., s_m\}$ and establish a mapping between each point o_i in **o** and each snippet in **s** in order since the outline contains sequential information. **In the fine-tuning stage**, we only calculate the loss from the **current** paired outline point and story snippet, while masking the other pairs. Hence, story partitioning may significantly impact the effectiveness of OC. We investigate two methods: **1) Average partition**, which divides stories by a constant number of sentences, and **2) Greedy Search**, which partitions stories by minimizing the distances between outline points and story snippets as well as the variance of the lengths of snippets. The objective function is as follows:

$$\mathcal{L}_{gs} = \alpha \sum_{i=1}^{m} \sum_{j \in s_i} dist(o_i, y_j) + \beta \sum_{i=1}^{m} (|s_i| - \mu_s)^2 \tag{2}$$

where $dist(\cdot)$ is the distance function, $|s_i|$ and μ_s denote the number of sentences in the story snippet s_i and the mean of $|s_i|_{i=1}^{m}$. α and β are weighting parameters. The second term of Eq. 2 is added to regularize the length distribution of story snippets. **In the inference stage**, we first use the prompt and o_1 to generate the first part. Then we iteratively generate each part based on the next outline sentence and the generated context. Specifically, for encoder-decoder models, we adjust the outline for the encoder input and the decoder prefix with context; whereas for decoder-only models, we concatenate each outline sentence and already generated parts as input. All generated parts are aggregated as the final result.

Zero-shot inference Similarly, we first use the prompt and o_1 with constraints including the expected length of the final story and the type (e.g., news article, story) to generate the first part. Then we iteratively generate each part based on the next outline sentence and the generated context. Finally, we aggregate all multi-round outputs. Since LLMs forget the previous settings in the multi-round interactions, we refine the aggregated output by reminding LLMs of the constraints and generation goal to obtain the final story. Details of prompts and examples are in Appendix A.2.

3.3 Unified Dual-task Learning

Summarization can be regarded as the inverse process of precise outline-conditioned generation. We propose a unified approach leveraging this task duality (denoted **Dual**) for both fine-tuning and zero-shot inference settings for precise outline-conditioned generation, to alleviate the imbalanced outline utilization issue and in turn improve generation quality.

Fine-tuning We introduce a dual task in the fine-tuning stage for the generation model θ_g, which summarizes the generated story with a model θ_s. We consider both generation by the summarization model, $P(\mathbf{o} \mid \mathbf{y}; \theta_s)$, and generation by the outline-conditioned generation model, $P(\mathbf{y} \mid \mathbf{o}; \theta_g)$, generally satisfy Eq.3 based on Bayes' rule. We omit the prompt $\mathbf{x}$ for simplicity.

$$P(\mathbf{o})P(\mathbf{y} \mid \mathbf{o}) = P(\mathbf{o}, \mathbf{y}) = P(\mathbf{y})P(\mathbf{o} \mid \mathbf{y}) \quad (3)$$

Inspired by Wei et al. (2019), we establish the bridge between the two tasks by adding a dual Lagrange regularizer term as Eq.4 to the standard fine-tuning loss. It aims to utilize the shared knowledge between the dual tasks to strengthen their connection by minimizing the difference between the model output and their approximate probability constraints.

$$\mathcal{L}_{\text{dual}} = \big[\log \hat{P}(\mathbf{o}) + \log P(\mathbf{y} \mid \mathbf{o}; \theta_g) \\ - \log \hat{P}(\mathbf{y}) - \log P(\mathbf{o} \mid \mathbf{y}; \theta_s) \big]^2 \quad (4)$$

where $\hat{P}(\mathbf{o}), \hat{P}(\mathbf{y})$ denote the marginal distributions of outlines and stories, respectively.

Zero-shot inference We propose to leverage the task duality to refine LLM zero-shot inference via in-context learning. Specifically, an LLM first generates a draft story based on the given outline. Then the model is required to summarize the generated draft. Based on the property of duality, the given outline and the generated summary should cover equivalent information. Therefore, the model is asked to compare the outline and the summary to reflect and further refine the generated text based on the discovered discrepancy. Importantly, we also investigate combining Dual with OC (denoted by **OC+Dual**) by conducting the summarization step in Dual on the story parts generated under OC. The details of the prompt and examples of OC+Dual are shown in Appendix A.3 and A.4.

4 Dataset Creation

To facilitate the study on the precise outline-conditioned generation task, we introduce two datasets, namely, WritingPrompt for Outline-conditioned Generation (**WPOG**) and CNN/DailyMail (**CDM**) (Hermann et al., 2015), from two distinct storytelling and news report scenarios, respectively.

WPOG is constructed based on the English WritingPrompt (WP) dataset (Fan et al., 2018), which includes human-written stories paired with writing prompts from an online forum. Since the original WP does not provide outlines, we randomly select 5% of the samples and use GPT-4[3] to generate the outlines for original stories. Moreover, we conduct the human evaluation to verify the reliability of generated outlines (the details and results are shown in Appendix D). The WPOG dataset is then built by pairing the generated outlines and the original writing prompts as input and the human-written stories as target references. To ensure fair comparisons between fine-tuning and zero-shot inference settings, we consider the limitations on the number of tokens output by BART and GPT-2 and exclude samples longer than 1024 tokens. We partition WPOG into 80% for training, 10% for validation, and 10% for testing.

CDM is an English dataset containing unique news stories (written by journalists with CNN and the Daily Mail) and the corresponding human-written highlights. For our task, the highlights are taken as outlines to represent the key plots of the news reports. We keep the original partition of training and validation sets as Nallapati et al. (2016), while sampling 10% from the original test set as our test set for reasonable turnaround time for zero-shot evaluations of LLMs. We follow the previous work (Fang et al., 2021) to obtain prompts and combine the prompt and outlines as input.

5 Experiments

The details of dataset preprocessing, dataset statistics, experimental setup, and implementations are presented in Appendix B.

---

[3]https://openai.com/research/gpt-4

5.1 Evaluation Metrics

For evaluations, we adopt three metrics that are commonly used for text generation tasks. **Rouge (R-n)** evaluates the n-gram recall between generated texts and human-written texts, where n can be 1, 2, or L (longest common subsequence) (Lin and Hovy, 2003). **BERTScore** computes the F1-score between generated texts and human-written texts, using BERT (Devlin et al., 2019) as a similarity measure (Zhang et al., 2019). **Distinct-4 (D-4)** measures the generation diversity (informativeness), based on the ratio of distinct 4-grams to all the generated 4-grams (Li et al., 2015).

We also propose three new metrics that measure how imbalanced the outline utilization is for precise outline-conditioned generation. **Distribution Variation (DV)** measures the difference between the distribution (D_j) of similarity between every outline sentence o_j and generated text $\mathbf{y}$, which is calculated as Eq. 5, where d_{KL} denotes KL-divergence between two distributions.

$$ DV = \frac{1}{|\mathbf{o}| \cdot (|\mathbf{o}| - 1)} \sum_{a,b \in \mathbf{o}, a \neq b} d_{KL}(D_a, D_b) $$

$$ (5) $$

Peak-value Distance (PD) indicates the distance of the most matched sentences ($P_j \in 1, 2, \ldots, k$) among the generated text $\mathbf{y}$ for every outline sentence o_j, respectively. It is defined as Eq.6.

$$ PD = \frac{1}{|\mathbf{o}| \cdot (|\mathbf{o}| - 1)} \sum_{a,b \in \mathbf{o}, a \neq b} |P_a - P_b| \quad (6) $$

Consistency Degree (CD) is proposed to assess outline utilization, motivated by the duality of summarization and precise outline-conditioned generation. Specifically, we use ChatGPT to summarize the generated text and contrast the summary with the original outline by computing Rouge-L of it against the outline as Consistency Degree (CD). The prompt used for CD is shown in Appendix A.5.

DV and PD could be considered as intrinsic metrics for outline utilization while CD is extrinsic. All of them are assessed from both faithfulness and structure perspectives (Section 1); and for each of them, a higher value suggests a more even outline utilization.

5.2 Automatic Evaluation

5.2.1 Main Results

Table 1 shows the main results of precise outline-conditioned generation on the two datasets CDM and WPOG under both fine-tuning and zero-shot inference settings. We report results based on the automatic evaluation metrics described in Section 5.1. For a fine-tuning baseline, we use the encoder-decoder BART-base[4]; for zero-shot inference baseline, we use ChatGPT-3.5-turbo[5]. For each baseline, we apply our proposed methods, **explicit outline utilization control (OC)**, **unified dual-task learning (Dual)**, and their combination **(OC+Dual)**.

In the fine-tuning setting, we find **both OC and Dual improve the baseline across all metrics on both datasets**, confirming the effectiveness of our methods. Notably, OC achieves a larger gain than Dual on CDM, especially on the diversity metric D-4, because the CDM outline reveals less information about the order and correspondence between points and news content. Since OC helps convey this information, it achieves larger gains than Dual on CDM. In contrast, OC has a smaller impact on WPOG than on CDM, because WPOG outlines are more explicit about the key plots and story elements. There is a clearer correspondence between paragraphs in the story and key points in the outline in WPOG, which likely reduces the benefit of OC. Notably, OC achieves substantial improvements on DV, PD, and CD on both CDM and WPOG, demonstrating the remarkable effectiveness of OC in facilitating more even outline utilization.

Combining OC and Dual (OC+Dual) improves the baseline, but it does not achieve the best results on most of the metrics compared to each individual approach. We think this is probably because of the conflict between the two methods when they are combined. To ensure that the dual loss from the summarization task can pass the relevant parts to the separate section in the generation under OC, we applied the same partitioning method to the summarization task, transforming it from a full-passage summarization to a partial point-wise summarization task, which may hurt the

---

[4]https://huggingface.co/facebook/bart-base
[5]https://chat.openai.com

6

Method	R-1	R-2	R-L	D-4	BERTScore	DV	PD	CD
CDM								
BART-base	38.79	10.86	16.26	79.47	84.62	2.21	1.09	14.74
w/ OC	**40.92**	11.82	**17.48**	87.94	86.76	**2.94**	**4.06**	**21.62**
w/ Dual	39.66	11.21	16.67	82.73	85.17	2.43	2.37	15.20
w/ OC + Dual	40.47	**12.04**	17.30	**89.26**	**87.21**	2.66	2.93	17.39
ChatGPT-3.5-turbo	41.17	12.17	17.62	97.46	84.97	2.63	3.44	18.96
w/ OC	41.69	12.78	18.67	**99.47**	85.48	2.75	4.46	**23.88**
w/ Dual	42.46	**13.10**	**19.06**	98.82	84.30	2.49	3.62	18.54
w/ OC + Dual	**42.90**	12.82	18.89	97.83	**87.29**	**2.85**	**4.76**	19.40
Ground Truth	-	-	-	89.24	-	3.16	8.67	29.62
WPOG								
BART-base	39.20	14.19	22.08	92.89	83.05	1.97	1.42	10.94
w/ OC	43.71	15.36	24.21	**96.30**	83.82	**2.76**	**3.94**	**19.63**
w/ Dual	**44.96**	**15.92**	**24.97**	93.24	86.12	2.61	2.85	15.97
w/ OC + Dual	42.77	14.89	23.74	95.71	**86.59**	2.49	3.29	17.68
ChatGPT-3.5-turbo	46.36	16.47	26.13	98.64	84.34	2.37	2.01	16.93
w/ OC	46.74	16.70	26.58	99.17	85.70	**2.86**	4.64	**20.11**
w/ Dual	47.42	17.12	27.30	99.25	**86.34**	2.71	3.85	19.07
w/ OC + Dual	**47.61**	**17.15**	**27.48**	**99.31**	86.22	2.85	**4.76**	19.40
Ground Truth	-	-	-	95.69	-	2.89	6.73	22.63

Table 1: Precise outline-conditioned generation performance of the base models and after applying our proposed **explicit outline utilization control (OC)**, **unified dual-task learning (Dual)**, and their combination **(OC+Dual)**. The best results for each metric in each group are in **bold**.

completeness of the summarization task.

In the zero-shot inference setting, **both OC and Dual substantially improve the baseline performance on both datasets, with Dual outperforming OC**. Compared with the baseline used in fine-tuning, LLMs have stronger power in contextual understanding. We believe this makes LLMs benefit more from restructuring and revising, rather than teaching them to generate segment by segment directly. The combination of OC+Dual further improves R-1 and PD metrics on both datasets and achieves comparable results on the other metrics, compared to using OC and Dual individually.

5.2.2 Analysis

Impact of model size and architecture for fine-tuning scenario. We compare the performance of encoder-decoder models BART-base (139M parameters) and BART-large (406M parameters) with decoder-only models GPT-2-base (124M parameters) and GPT-2-large (774M parameters) for fine-tuning under outline-conditioned generation, as shown in Table 3 (Appendix C). Both OC and Dual improve all the baseline models regardless of their model size and architecture. When comparing OC and Dual, we observe that BART-base and GPT-2-base with OC achieve better results than Dual on all metrics except CD. This ob-

servation highlights that both encoder-decoder and decoder-only models exhibit an imbalanced outline utilization issue, and it underscores the effectiveness of OC in mitigating this concern.

Impact of different LLMs for zero-shot inference scenario. We compare the performance of Vicuna-13B and ChatGPT for zero-shot inference under precise outline-conditional generation, as shown in Table 4 (Appendix C). The results show that **both OC and Dual improve** the performance of both baseline models on all metrics and datasets. This indicates that **our methods are effective** for different kinds of LLMs under zero-shot inference scenarios. Unlike ChatGPT, Vicuna-13B benefits more from OC on R-1 and R-L, which suggests that OC works better on smaller models.

Impact of partition methods for outline control in fine-tuning scenario. We propose two ways (average partition and greedy search) to make partitions for stories, as mentioned in section 3. We compare the performance of these two ways on BART-base and GPT-2-base for fine-tuning scenarios. The results are shown in Figure 4 (Appendix C). The performance of greedy search **significantly outperforms** the average partition, which indicates that the suitable partition contributes to the generation performance by establishing the mapping between outlines and parts

7

of stories. However, the different partitioning methods show consistent improvements in outline-related metrics (i.e., DV, PD, and CD) compared to the baseline model. This demonstrates the effectiveness of outline control in mitigating the problem of imbalanced utilization of the outline information.

5.3 Human Evaluation

Besides the validation of the WPOG dataset, we conduct another two human evaluation experiments to answer the following research questions:

RQ1: How is the text generation quality of the models evaluated in this task in two different scenarios?

RQ2: Can the new metric proposed in this paper correctly reflect the usage of outlines?

In this task, we sample 20 outlines and compare corresponding stories generated by the base and OC+Dual variants in both fine-tuning and zero-shot inference. We ask ten evaluators to rate the quality of stories in different categories: fluency, coherence and flow, completeness, relevance, beginning, closure, and outline ordering. The details about criteria and questions are shown in Figure 9 (Appendix F).

In Figure 7 (Appendix C), we show the overall score and detailed performance of the four methods. It is observed that variants employing OC+Dual strategies have shown improvements compared to the base model, which is consistent with the trend reflected by automatic evaluation results. It is notable that, for fine-tuning inference, the variant gets strong improvement on two structure metrics, i.e., closure and outline ordering. It demonstrates that the strategies we proposed can alleviate the issue that the outline information is overused at the beginning. In terms of faithfulness, the variants achieve 0.5 and 0.18 points higher than the base model in the combination of completeness and relevance over fine-tuning and zero-shot inference, respectively.

For RQ2, we compare the results of our proposed outline usage automatic metrics: Peak-value Distance (PD), Distribution Variation (DV), and Consistency Degree (CD), and four outline-relevant human evaluation metrics: completeness, relevance, closure, and outline ordering as shown in Figure 3. In overall comparison, they show similar trends in distin-
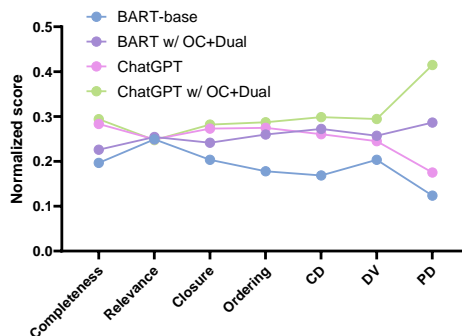

Figure 3: Comparison among human evaluation and automatic metrics.

guishing the base model and their corresponding variants. The result on extrinsic metric CD is close to outline ordering, which demonstrates the way to reconstruct the outline could reflect the structure character of generated text. When it comes to intrinsic metrics, the variation in the distribution of outline similarity demonstrates greater robustness in comparison to the distance between sentences that are most similar to the outline. This robustness aligns better with human evaluation metrics.

6 Conclusion

In this paper, we introduce a novel text generation task called precise outline-conditioned generation, which generates stories based on specific, sentence-level outlines. We construct two new benchmark datasets, WPOG and CDM, for this task. We provide strong baselines based on fine-tuning models such as BART and GPT-2, and evaluating zero-shot performance of models such as ChatGPT and Vicuna. We identify an issue related to the imbalanced utilization of the outline information in the precise outline-conditioned generation, which affects both fine-tuning and zero-shot models. To address this issue, we have proposed a unified explicit outline utilization control approach and a novel framework that leverages the task duality between summarization and generation. We have shown that our proposed approaches effectively alleviate the issue of imbalanced outline utilization and enhance the quality of precise outline-conditioned text generation for both fine-tuning and zero-shot settings. We hope that our work will inspire more research on this challenging and interesting task.

8

7 Limitations

As discussed in this paper, the proposed task of precise outline-conditioned generation could be regarded as a key subtask in conditional or open-ended long-text generation. However, we only evaluate our methods on texts with a length limitation of one thousand tokens (approximately 750 words). We do not assess the performance of our strategies on longer texts (thousands of words or even longer) since the quality of long text generation heavily depends on human evaluation, which is time-consuming and expensive.

8 Ethical Considerations

Our work focuses on advancing outline-conditioned text generation technology. We acknowledge that similar technology can be utilized to generate deceptive or manipulative content such as fake news. Our research provides unified strategies that are utilized by base language models, compatible with their detection and restriction of harmful information and false information. Additionally, the result demonstrates our method could improve the faithfulness of generated text to the given outline, making it easy for further detection and control over the input outline rather than whole articles. The zero-shot inference prompts are currently tailored exclusively for the English language, and extending them to other languages would necessitate further adjustments. The effectiveness of our methods could potentially be compromised in languages with fewer resources, as we rely extensively on pre-trained language models that may exhibit diminished performance in such linguistic contexts.

References

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2018. Yake! collection-independent automatic keyword extractor. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR*, volume 10772 of *Lecture Notes in Computer Science*, pages 806–810. Springer.

Fredrik Carlsson, Joey Öhman, Fangyu Liu, Severine Verlinden, Joakim Nivre, and Magnus Sahlgren. 2022. Fine-grained controllable text generation using non-residual prompting. In *Annual Meeting of the Association for Computational Linguistics*.

Hong Chen, Raphael Shu, Hiroya Takamura, and Hideki Nakayama. 2021. Graphplan: Story generation by planning with event graph. In *INLG*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *ArXiv*, abs/1912.02164.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. *ArXiv*, abs/1902.01109.

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. *arXiv preprint arXiv:2101.00822*.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph M. Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Conference on Empirical Methods in Natural Language Processing*.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. A probabilistic formulation of unsupervised text style transfer. *ArXiv*, abs/2002.03912.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *ArXiv*, abs/1506.03340.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B. Dolan. 2015. A diversity-promoting objective function for neural conversation models. *ArXiv*, abs/1510.03055.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pages 150–157.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. Plotmachines: Outline-conditioned generation with dynamic plot state tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295.

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2020. Investigating pretrained language models for graph-to-text generation. *ArXiv*, abs/2007.08426.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, pages 1–20.

Zhihong Shao, Minlie Huang, Jiangtao Wen, Wenfei Xu, and Xiaoyan Zhu. 2019. Long and diverse text generation with planning-based hierarchical variational model. In *EMNLP*.

Yixuan Su, David Vandyke, Sihui Wang, Yi-mai Fang, and Nigel Collier. 2021. Plan-then-generate: Controlled data-to-text generation via planning. In *Conference on Empirical Methods in Natural Language Processing*.

Xiaofei Sun, Zijun Sun, Yuxian Meng, Jiwei Li, and Chun Fan. 2022. Summarize, outline, and elaborate: Long-text generation via hierarchical supervision from extractive summaries. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6392–6402.

Pradyumna Tambwekar, Murtaza Dhuliawala, Lara J. Martin, Animesh Mehta, Brent Harrison, and Mark O. Riedl. 2018. Controllable neural story plot generation via reward shaping. In *International Joint Conference on Artificial Intelligence*.

Bowen Tan, Zichao Yang, Maruan Al-Shedivat, Eric P. Xing, and Zhiting Hu. 2020. Progressive generation of long text. *ArXiv*, abs/2006.15720.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Neural Information Processing Systems*.

Rose E. Wang, Esin Durmus, Noah D. Goodman, and Tatsunori Hashimoto. 2022. Language modeling via stochastic processes. *ArXiv*, abs/2203.11370.

Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Topic-guided variational auto-encoder for text generation. In *North American Chapter of the Association for Computational Linguistics*.

Bolin Wei, Ge Li, Xin Xia, Zhiyi Fu, and Zhi Jin. 2019. Code generation as a dual task of code summarization. *Advances in neural information processing systems*, 32.

Jingjing Xu, Xuancheng Ren, Y. Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. 2018. A skeleton-based model for promoting coherence among sentences in narrative story generation. *ArXiv*, abs/1808.06945.

Kevin Yang and Dan Klein. 2021. Fudge: Controlled text generation with future discriminators. *ArXiv*, abs/2104.05218.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. Doc: Improving long story coherence with detailed outline control. In *Annual Meeting of the Association for Computational Linguistics*.

Kevin Yang, Nanyun Peng, Yuandong Tian, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Conference on Empirical Methods in Natural Language Processing*.

Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. *ArXiv*, abs/1811.05701.

Hanqing Zhang and Dawei Song. 2022. Discup: Discriminator cooperative unlikelihood prompt-tuning for controllable text generation. In *Conference on Empirical Methods in Natural Language Processing*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Yizhe Zhang, Guoyin Wang, Chunyuan Li, Zhe Gan, Chris Brockett, and Bill Dolan. 2020. Pointer: Constrained progressive text generation via insertion-based generative pre-training. In *Conference on Empirical Methods in Natural Language Processing*.

A Prompt Design

Here we list the prompt we used under the zero-shot scenario for reproducing our experiment results and the one we used in evaluating consistent degree (CD). Here we take the prompt on CDM as example, while WPOG only replaces the "news report" by "story". It is noticeable that both ChatGPT and Vicuna are sharing these prompts.

A.1 Prompt - Baseline

User: Please generate a [XXX] words news report with given first sentence and outlines. The first sentence is: [XXX]. And the outline of this news report is [XXX].

Agent: [XXX].

A.2 Prompt - Explicit Outline Utilization Control (OC)

User: Your final goal is to generate a [XXX] words news report based on given first sentence and outlines. The first sentence is: [XXX] And the outline of this news report is: [XXX]. Firstly, you could generate a part of the news report corresponding to following plot [XXX].

Agent: [XXX].

User: Continue to generate a part of the news report followed by your previous output while the plot is corresponding to the following plot [XXX].

Agent & **User**: ... repeat m-1 times

User: Now connect all the paragraphs you've written and polish them to a [XXX] tokens essay to achieve the final goal to generate an around [XXX] tokens news report with given first sentence and outlines. First sentence: [XXX] Outline: [XXX].

Agent: [XXX].

A.3 Prompt - Unified Dual-task Learning (Dual)

User: Please generate a [XXX] words news report with given first sentence and outlines. The first sentence is: [XXX]. And the outline of this news report is [XXX].

Agent: [XXX].

User: Please summarize it into [XXX] one-sentence points.

Agent: [XXX].

User: Please compare with the following true outline and rethinking how to improve the quality of outline-conditioned news report generation: [XXX].

Agent: [XXX].

User: Based on the knowledge you just learned, regenerate a [XXX] words news report with given first sentence and outlines. The first sentence is: [XXX]. And the outline of this news report is [XXX].

A.4 Prompt - Combo: OC + Dual

*** Repeat all steps in OC mode ***
User: Please generate a [xxx] tokens story with the given first sentence and outlines. The first sentence is: [XXX]. And the outline of this story is [XXX]

Agent: [XXX].

User: Please summarize it into an outline that has [XXX] one-sentence points.

Agent: [XXX].

User: Please compare with the following true outline and rethinking how to improve the quality of outline-conditioned news report generation. True outline: [XXX].

Agent: [XXX].

User: Based on the knowledge you just learned, regenerate a [XXX] words news report with given first sentence and outlines.

The first sentence is: [XXX]. And the outline of this news is [XXX].

User: Refine your output to a [XXX] words news report.

A.5 Prompt - Consistency Degree (CD)

User: Please generate an outline for the given news report. The news report: [XXX]

Agent: [XXX].

B Experimental Setup

Dataset Preprocessing In the data preprocessing, we use the first sentence as input. While CNN/DailyMail usually includes some meaningless openings for news reports (e.g., reporter's name), we made an additional filter for this issue, which only accepts openings with more than 7 words. Moreover, we aim to generate long text involving abundant semantic and structural information. Here we filter out all cases with less than 64 words. We cut off 40 sentences to test model performance.

Dataset Statistics Table 2 provides the statistics of CDM and WPOG. Notably, WPOG has much longer outlines than CDM while CDM comprises substantially more training samples than WPOG.

Table 2: Statistics of the CDM and WPOG datasets. Avg. length denotes the number of tokens from the BART tokenizer (Lewis et al., 2020).

Dataset	CDM	WPOG
# Train	287,113	6,982
# Valid	13,378	866
# Test	1,149	866
Avg. length of story	780.62	542.57
Avg. length of outline	56.20	165.12

Implementation Details Our code implementations are mainly based on Pytorch 2.0 and the Huggingface library. All experiments including model finetuning (e.g., BART, GPT-2) and zero-shot inference (e.g., Vicuna) are running on 8 NVIDIA V100 32G GPUs. The total computation cost is about 1500 GPU hours. For fine-tuning BART and GPT-2, we set the batch size as 4, and training epochs as 3 on CDM, and 20 on WPOG since they

have different scales of training set. The hyperparameters α, β in outline control are set to 0.8, and 0.05, which are obtained by grid search from $\{0.5, 0.6, 0.7, 0.8, 0.9\}$ and $\{0.005, 0.01, 0.05, 0.1\}$. We report the mean value for all results after running the experiments 3 times. It is noticeable that the structure of GPT-2 is different from BART, we concatenate both the opening and outline together through prompts as input. For zero-shot scenarios, we use OpenAI's 3.5-turbo API and the Vicuna-13B [6] model to achieve it. For dataset creation, we use GPT-4[7] API to generate the outlines. According to the official price and our usage, we spend around $700 on ChatGPT API costs.

For human evaluation, we recruited 10 graduate students from the university as volunteers. They are from the United States, China, and India. We provide them with $5 for dataset verification and $15 for outline-conditioned generation evaluation.

C More Analysis Results

Table 3 demonstrates the impact of model sizes and architectures for fine-tuning LMs for precise outline-conditioned generation on the CDM dataset. Table 4 shows the impact of using different LLMs under zero-shot inference setting for precise outline-conditioned generation on the CDM dataset. Figure 4 illustrates the impact of different partition methods on the performance of our proposed **explicit outline utilization control (OC)** on CDM. Figure 7 shows the overall score and detailed performance of human evaluation for the four methods (i.e., ChatGPT, ChatGPT with OC+Dual, BART, BART with OC+Dual).

Case study about imbalanced outline information usage The similarity measurement between the outline and the entirety of the generated text provides a clear illustration of this issue, as the case study depicted in Figure 5. In the text generated by BART (Figure 5(b)), the utilization of information is primarily focused at the beginning, contrasting with the distribution observed in the ground truth (Figure 5(a)). Furthermore, a parallel trend is observed in large language models (LLMs),

---

[6]https://huggingface.co/lmsys/vicuna-13b-v1.5
[7]https://openai.com/gpt-4

12

Method	R-1	R-2	R-L	DV	PD	CD
BART-base	38.79	10.86	16.26	2.21	1.09	14.74
w/ OC	**40.92**	**11.82**	**17.48**	**2.94**	**4.06**	**21.62**
w/ Dual	39.66	11.21	16.67	2.43	2.37	15.20
BART-large	40.26	10.75	17.05	**2.51**	2.78	16.28
w/ OC	**41.39**	**12.06**	**18.43**	2.46	**3.02**	**23.81**
w/ Dual	40.41	11.69	17.74	2.36	2.96	19.46
GPT-2-base	34.63	8.29	13.48	1.92	1.21	9.76
w/ OC	**36.92**	**10.22**	**15.53**	**2.64**	**3.39**	15.94
w/ Dual	36.48	10.19	15.19	2.30	2.90	**16.24**
GPT-2-large	37.95	9.30	15.86	2.35	1.74	12.57
w/ OC	38.20	10.25	16.14	**2.76**	**3.42**	14.09
w/ Dual	**38.87**	**10.69**	**16.49**	2.49	2.19	**15.78**

Table 3: Impact of model sizes and architectures of fine-tuning LMs for precise outline-conditioned generation on CDM. The best results for each metric in each group are in **bold**.

Method	R-1	R-2	R-L	DV	PD	CD
ChatGPT	41.17	12.17	17.62	2.63	3.44	18.96
w/ OC	41.69	12.78	18.67	**2.75**	**4.46**	**23.88**
w/ Dual	**42.46**	**13.10**	**19.06**	2.49	3.62	18.54
Vicuna-13B	39.43	11.28	17.92	2.14	2.90	16.83
w/ OC	**41.26**	11.74	**18.20**	2.36	3.17	**19.07**
w/ Dual	40.90	**12.42**	18.07	**2.42**	**3.29**	18.29

Table 4: Impact of using different LLMs under zero-shot inference setting for precise outline-conditioned generation on CDM. The best results for each metric in each group are in **bold**.

such as ChatGPT (Figure 5(c)), signifying that the imbalanced utilization of outlines is a pervasive issue in text generation. The experimental results, depicted in Figure 5(d), validate the efficacy of this approach.

D   Dataset Quality Validation

We evaluate the quality of AI-annotated outlines by accuracy, completeness, and clarity. The definition and criteria are shown in Figure 8 (Appendix F). Specifically, we randomly select 20 pairs of <story, outline> from WPOG and another 20 pairs from CDM. Each evaluator is asked to rate from 1 to 5 for a sample of 5 stories from each of the two datasets. To avoid bias, we ensure that each sample is rated by two different evaluators.

As Figure 6 illustrates, the WPOG's outline achieves higher scores in both completeness and clarity. while it slightly lags behind CDM in terms of accuracy. We believe this is due to the fact that the model adds some details to maintain the structural integrity of the outlines, thus sacrificing some accuracy. Taking these
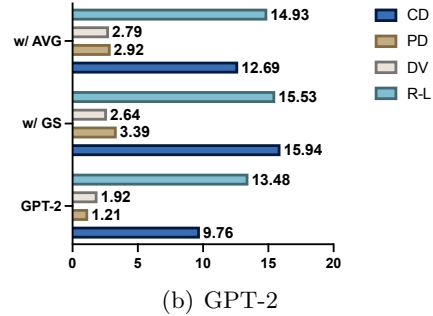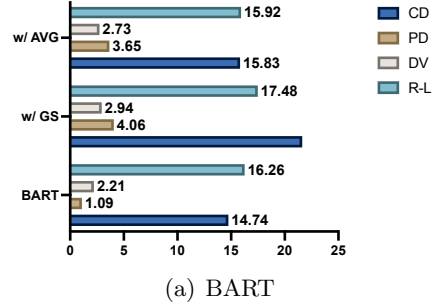

(a) BART

(b) GPT-2

Figure 4: The impact of different partition methods on the performance of **explicit outline utilization control (OC)** on CDM.

three metrics into account, the quality of the WPOG dataset is compelling.

E   Examples of Precise Outline-conditioned Generation

Table 5 shows an example of how the baseline models of BART and ChatGPT generate stories based on prompts and outlines.

F   Questionnaire for Human Evaluation

The questionnaire for dataset validation and quality of generated stories consists of explanations and questions, as shown in Figure 8 and Figure 9, respectively.

Examples of precise outline-conditioned generation:	
Outline	Tourists and locals queue for several hours to get their hands on Jenny's butter cookies .
	People are even hired to stand in line to buy the cookies, which are later sold at an up-to-70% mark-up
	Food frenzies have also taken place in other parts of the world .
Prompt	There's a booming black market in Hong Kong, but it's not for fake Apple Watches, or the iPhone.
Output (BART)	There's a booming black market in Hong Kong, but it's not for fake Apple Watches or iPhones. Tourists and locals queue for hours to buy Jenny's butter cookies. People are even hired to stand in line and later sell the cookies at a hefty markup, sometimes as much as 70%. These food frenzies aren't limited to Hong Kong; they're happening worldwide. From limited-edition Kit Kats in Japan to bagel crazes in New York, food tourism is influencing travel. This trend raises questions about what drives such desires for specific foods and what it means for the culinary and tourism industries. Jenny's butter cookies remain a testament to the power of taste and the lengths people go for a delicious treat.
	Jenny's butter cookies, with their irresistible allure, serve as a sweet testament to the lengths people will go to for a bite of something truly special. These worldwide culinary obsessions beg the question: What fuels these fervent desires for specific foods? The answer may lie in the power of taste and the universal appeal of exceptional culinary experiences. Once secured, these cookies are later resold at an astonishing markup, sometimes reaching up to 70% above the original price. It's a business strategy that has tapped into the undeniable allure of Jenny's cookies and the willingness of customers to pay a premium for them. Queueing for hours has become a common sight as both locals and tourists are eager to savor these delectable treats. But this craze doesn't stop at the cookie shop's door.
Output (Chat-GPT):	There's a booming black market in Hong Kong, but it's not for fake Apple Watches or iPhones. The latest craze sweeping the vibrant city isn't about tech gadgets; it's about butter cookies. Jenny's butter cookies, to be precise, have become a culinary sensation, with both locals and tourists eagerly joining long queues for a taste of these delectable treats. The scene is surreal. People stand in line for several hours, determined to get their hands on Jenny's butter cookies. The allure of these sweets is irresistible, with their reputation for unmatched flavor and melt-in-your-mouth goodness. Tourists visiting Hong Kong are often captivated by the buzz surrounding these cookies, making them an essential item on their travel itineraries. But Hong Kong isn't the only place where food frenzies have taken hold.
	However, what's intriguing is how this intense demand has given birth to a clandestine world of commerce. To meet the soaring demand for these cookies, some individuals are hired to stand in line and purchase the coveted treats, which are later sold at a staggering markup. Reports suggest that these cookies, originally priced reasonably, are now reselling for up to a 70% increase in value. It's a clear reflection of the extent to which people are willing to go for a taste of this newfound delicacy.
	Across the globe, similar phenomena have unfolded. From Taiwan's famous pineapple cakes to Parisian macarons, food enthusiasts are demonstrating an insatiable appetite for unique and sought-after culinary experiences, prompting culinary tourism to flourish.
	As the butter cookie craze continues to captivate hearts and taste buds, it raises questions about the nature of trends in the modern world. What drives the fervent desire for specific foods, and what do these food frenzies mean for the culinary industry and global tourism? In the midst of it all, Jenny's butter cookies remain a sweet testament to the power of taste, the lengths people will go for a delicious bite, and the intriguing world of culinary commerce.

Table 5: Examples of precise outline-conditioned generation over CDM. The highlights reflect the usage of corresponding outline information
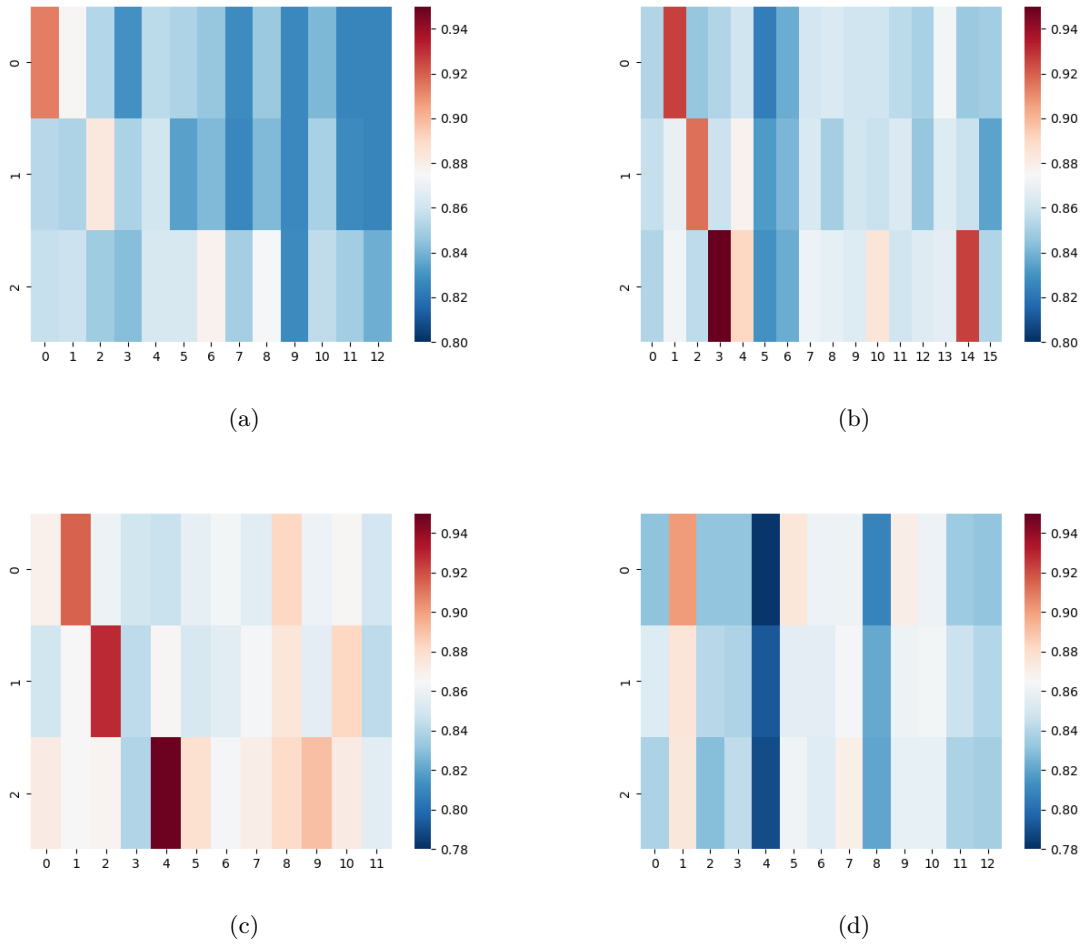
Figure 5: The case study of outline similarity within the entire story in CDM. (a) Ground truth text, (b) Generated text from BART-base without outline control, (c) Generated text from ChatGPT (gpt-3.5-turbo) without outline control, (d) Generated text from BART-base with outline control. The Y-axis represents the outline bullet points, while the X-axis represents the sentences in the generated text. The values indicate the corresponding similarity between the outline bullet point and the generated sentences.
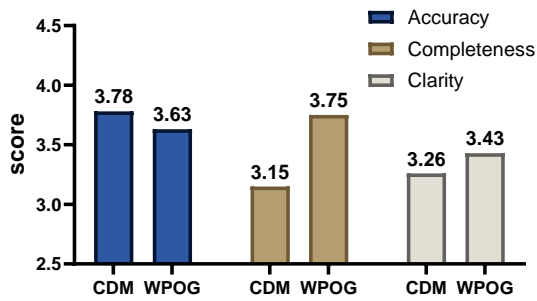
Figure 6: The human evaluation on dataset quality validation.
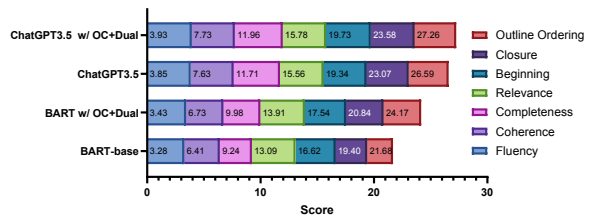
Figure 7: The human evaluation on overall score and detailed performance of the four methods.

Dataset Evaluation

In this test, you will read 10 articles and their corresponding outlines. Please rating the **quality of outlines** from 1 (lowest) to 5 (highest) based on following criteria:

1. **Accuracy**:
   - Does the summary accurately represent the main events, characters, and plot of the original story?
   - Are there any factual inaccuracies or misrepresentations?
2. **Completeness & Coverage**:
   - Does the outlines cover essential elements of the story, including the beginning, middle, and end?
   - Are critical plot points and character developments included?
3. **Clarity and Coherence**:
   - Is the summary written in a clear and coherent manner, making it easy for the reader to follow the storyline?
   - Are transitions between different parts of the summary smooth and logical?

Please notice that your task is to evaluate the quality of outlines not the articles.

(a)

	1: Poor	2: Mediocre	3: Acceptable	4: Satisfactory	5: Excellent
Accuracy	○	○	○	○	○
Completeness	○	○	○	○	○
Clarity	○	○	○	○	○

(b)

Figure 8: An example about explanation (a) and questions (b) in the questionnaire of dataset validation from human evaluation.

Outline-based generation Performance

Please evaluate the following stories according to the given four story outlines. It is evaluated in following aspects:

Fluency & Coherence:

1. **Fluency:** Evaluate how well the generated text flows naturally and is free from grammatical errors, awkward phrasing, or language issues.
2. **Coherence and Flow:** Check for logical connections between ideas and smooth flow of ideas. Highlight parts of the essay that appear disjointed or lack clear organization.

Completeness & Relevance

1. **Completeness:** Check that the generated text covers all the points and sub-points of the overview. Assess whether any key information is missing from the generated text. Ensure that the text responds fully to the outline.
2. **Relevance**: Assess the inclusion of relevant and supporting details. Ensure that the text does not include irrelevant or extraneous information.

Structure and Outline Usage

1. **Beginning:** Evaluate how effectively the story begins and sets the stage for the narrative. Is the opening compatible with the first sentence.
2. **Closure:** how the story concludes and whether it provides a satisfying closure to the narrative.
3. **Outline Ordering:** Outline ordering evaluates how well the generated story follows a structured outline or plan. It assesses whether the story elements are presented in a logical sequence.

(a)

Fluency and Coherence *

	1: Poor	2: Mediocre	3: Acceptable	4: Satisfactory	5: Excellent
Fluency	○	○	○	○	○
Coherence	○	○	○	○	○

Completeness and Relevance *

	1: Poor	2: Mediocre	3: Acceptable	4: Satisfactory	5: Excellent
Completeness	○	○	○	○	○
Relevance	○	○	○	○	○

Structure *

	1: Poor	2: Mediocre	3: Acceptable	4: Satisfactory	5: Excellent
Beginning	○	○	○	○	○
Closure	○	○	○	○	○
Outline Ordering	○	○	○	○	○

(b)

Figure 9: An example about explanation (a) and questions (b) in the questionnaire of the quality of generated stories from human evaluation.