Parameter-Efficient Style-Controlled Summarization: An Investigation into LoRA Dynamics and Stylistic Collapse

Josemaria Louis Fernando
School of Sciences, Engineering, & Technology
University of Asia and the Pacific
Pasig City, Philippines
josemaria.fernando@uap.asia

Elmer Peramo
Advanced Science and Technology Institute
Department of Science and Technology
Quezon City, Philippines
elmer@asti.dost.gov.ph

Dantenelo Santi Barretto
School of Sciences, Engineering, & Technology
University of Asia and the Pacific
Pasig City, Philippines
dantenelosanti.barretto@uap.asia

Ashley Trish Soriano
School of Sciences, Engineering, & Technology
University of Asia and the Pacific
Pasig City, Philippines
ashleytrish.soriano@uap.asia

***Abstract*—Abstractive summarization with large pre-trained Transformer models struggles to reliably control stylistic attributes (e.g., "punchy" vs. "neutral" tone) without sacrificing content fidelity, especially under parameter-efficient adaptation. This work systematically examines Low-Rank Adaptation (LoRA) on PEGASUS-Large for style-conditioned headline generation using special control tokens and an ablation over eight configurations that vary LoRA rank (r $\in \{8, 16, 32, 64\}$) and target modules (Attention-Only vs. FFN-Expanded).**

Experiments reveal a phenomenon we term *Stylistic Collapse*, where the model's strong extractive bias overwhelms the parameter-efficient style signal. Across all configurations, the Identical Output Rate remains high (above 62%), with no consistent improvement from increased LoRA capacity, even though factual consistency (entailment > 0.85) and content fidelity (ROUGE-1 ≈ 0.52 for neutral) remain strong. These results suggest that, in our PEGASUS-Large setting and dataset, standard parameter-efficient fine-tuning may be insufficient for enforcing subtle stylistic control, and motivates future exploration of objectives (e.g., contrastive losses or reinforcement learning) that explicitly reward stylistic divergence from the base model's inductive bias.

***Index Terms*—Abstractive Summarization, LoRA, Style Transfer, PEGASUS, Natural Language Processing.**

I. Introduction

Abstractive summarization models such as PEGASUS and BART have demonstrated strong performance across news and document summarization tasks. However, these models are typically optimized for a single, neutral summarization objective, making it unclear whether they can be adapted to reliably express stylistic variation such as "neutral" versus "punchy" headlines. Parameter-efficient fine-tuning (PEFT) methods such as LoRA offer lightweight adaptation, but their effectiveness for style-controlled summarization, where subtle stylistic changes must coexist with strict content preservation, remains uncertain. Early work on neural style control showed that conditioning language models on discrete stylistic attributes can steer properties such as sentiment and descriptiveness [1], and subsequent studies have explored deep learning approaches to text style transfer more broadly [2].

In this work, we investigate whether LoRA-based fine-tuning is sufficient for controllable headline generation under realistic constraints: limited annotated data, subtle human-defined stylistic distinctions, and a strong extractive bias in the base model. Rather than optimizing for the best-performing system, our objective is to provide a diagnostic (negative-result) analysis of failure modes that arise when applying parameter-efficient methods to this task. Using a dataset of Philippine news snippets annotated with neutral and punchy headlines, we introduce and systematically vary LoRA configurations that expand adapters to both attention and feed-forward layers. The findings reveal a consistent phenomenon of *stylistic collapse*, in which PEGASUS-Large largely ignores style tokens and generates nearly identical summaries regardless of the prompt, highlighting architectural and data-related factors that drive this collapse.

II. Related Work

A. Abstractive Summarization

Transformer-based encoder–decoder models have become the dominant architecture for abstractive summarization. Among these, PEGASUS introduces a pre-training strategy tailored to summarization through gap-sentence generation, where important sentences are masked and predicted from the remaining context [3]. Unlike generic masked language

*This work was submitted to the International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2025.

modeling, this objective closely matches the downstream summarization task, and PEGASUS achieves strong ROUGE performance across multiple domains, including news, scientific writing, and government documents [3]. Building on such models, recent work has focused on parameter-efficient fine-tuning methods such as LoRA [4] and QLoRA [5] to adapt large models with limited additional parameters; see [6] for a comprehensive survey of PEFT techniques. More broadly, our work fits into the line of controllable text generation with transformer-based pre-trained language models, where style, sentiment, or other attributes are manipulated at training or inference time [7].

B. Style Transfer

Text style transfer (TST) aims to control attributes such as formality, politeness, emotion, and ideology while preserving the underlying semantic content. A recent survey by Jin et al. reviews more than one hundred neural TST methods and shows that most approaches rely on parallel corpora, adversarial learning, or disentangled latent representations [2]. Although these methods demonstrate stylistic control, they frequently suffer from a trade-off between style strength and content preservation, especially in low-resource or non-parallel settings.

Within the specific domain of headline generation, Jin et al. introduced the task of Stylistic Headline Generation (SHG) and proposed the TitleStylist multitask framework to generate headlines with explicit styles such as humor, romance, and clickbait [8]. Their model jointly learns summarization and style reconstruction through parameter sharing, producing headlines that outperform both neural baselines and human references in terms of perceived attractiveness. However, this approach requires full-model training and complex multi-objective optimization, making it computationally expensive.

An alternative line of work explores inference-time style control without retraining. Cao and Wang proposed modifying decoder hidden states using external style classifiers or constraining lexical choices during decoding to impose stylistic control [9]. Although this approach allows flexible post-hoc style manipulation, it depends heavily on high-quality external classifiers and does not alter the internal style representations learned by the model.

C. Parameter-efficient Fine-tuning with LoRA and QLoRA

As model sizes increase, full fine-tuning becomes increasingly impractical. Hu et al. introduced Low-Rank Adaptation (LoRA), which freezes the original model parameters and injects trainable low-rank matrices into each linear projection layer [4]. This reduces the number of trainable parameters by several orders of magnitude while maintaining performance comparable to full fine-tuning across models such as GPT-2, GPT-3, RoBERTa, and DeBERTa [4].

Dettmers et al. extended this approach with QLoRA, which combines 4-bit quantization with LoRA adapters to enable the fine-tuning of models with up to 65 billion parameters on a single GPU [5]. Their method introduces NormalFloat (NF4) quantization, double quantization, and paged optimizers to drastically reduce memory consumption without sacrificing performance. Importantly, they show that adapting not only attention layers but also feed-forward network (FFN) layers is critical for tasks that require deeper semantic transformation [5].

Despite these advances, most applications of LoRA and QLoRA focus on domain adaptation, instruction tuning, and dialogue systems. Their effectiveness for style-conditioned summarization, especially when stylistic differences are subtle (e.g., neutral vs. punchy headlines), remains underexplored.

D. Positioning of the Present Study

Prior work shows that PEGASUS is a strong abstractive summarization backbone [3], that stylistic headline generation is feasible under full-model multitask training [8], and that LoRA-based fine-tuning enables efficient adaptation of large language models [4], [5]. However, to the best of our knowledge, no existing work systematically examines whether parameter-efficient fine-tuning alone is sufficient for reliable style control in summarization or whether such setups exhibit a tendency toward stylistic collapse, where distinct style prompts produce near-identical outputs. This study directly targets this gap.

III. Methodology

This study employed a structured experimental pipeline to investigate whether Low-Rank Adaptation (LoRA) can provide reliable stylistic control in abstractive headline generation when applied to PEGASUS-Large. The methodology was intentionally designed to test the feasibility and limitations of parameter-efficient fine-tuning for style transfer under realistic constraints, including limited parallel data, subtle stylistic distinctions, and the strong extractive bias inherent in PEGASUS.

A. Dataset Preparation

A dataset of short Philippine news snippets (1–3 sentences) was curated for the study. Each snippet was paired with two human-written headline variants reflecting distinct stylistic intents:

- **Neutral** — concise, factual, and stylistically plain.
- **Punchy** — attention-grabbing, utilizing strong verbs or emotionally loaded phrasing.

To prevent information leakage across splits, all news items were partitioned at the article level into 70% training, 15% validation, and 15% test. The PEGASUS tokenizer was extended to include the control tokens `<neutral>` and `<punchy>`, with the embedding matrix resized to ensure these stylistic markers acquired learnable vector representations.

1) Dataset Statistics: The final corpus contains 421 unique news items (1–3 sentence snippets), each paired with one neutral and one punchy headline. Data were split at the article level into 70% train (294 items), 15% validation (63), and 15% test (64) to prevent story leakage.

2) Annotation Guidelines:

- Neutral: concise factual headline, mirrors the key entity/action in the snippet, avoids emotional language, no punctuation intended for effect (no exclamation marks or trailing colon patterns unless present in source). Example style: "Town councillor dies in road crash in Zamboanga."
- Punchy: attention-grabbing phrasing, uses strong verbs and emotive adjectives; often includes marker punctuation such as ":" or "!" followed by a short dramatic tag, e.g., "Tragedy strikes: Town councilor killed in Zamboanga!"
- Punchy candidates were initially drafted by an instruction-tuned large language model and then manually curated and edited by human annotators to ensure readability and relevance.

B. Model Architecture and LoRA Configuration

PEGASUS-Large (568M parameters) was selected as the base model due to its strong summarization performance and gap-sentence generation pre-training objective. However, PEGASUS is known to exhibit a strong extractive bias, making it an ideal test bed for assessing whether PEFT methods can overcome stylistic rigidity.

We evaluated eight LoRA variants spanning ranks $r \in \{8, 16, 32, 64\}$ and two target-module sets. The scaling factor α was fixed to $2r$ for all runs. Attention-Only (A-Only) adapters modified the query and value projections, while FFN-Expanded (FFN-E) additionally targeted key and output projections and both feed-forward layers, following evidence that FFN adaptation benefits tasks requiring deeper semantic transformation [5].

We defined two primary sets of target modules for the LoRA adapters:

1) **Attention-Only (A-Only):** A standard setup where adapters are applied only to the core attention mechanism projections, as illustrated in Fig. 1.
2) **FFN-Expanded (FFN-E):** An aggressive expansion, incorporating adapters into the Feed-Forward Network (FFN) layers to maximize adaptation capacity and facilitate the capture of semantic stylistic features, as suggested by prior work [5]. The FFN-E configuration is visually represented in Fig. 2.

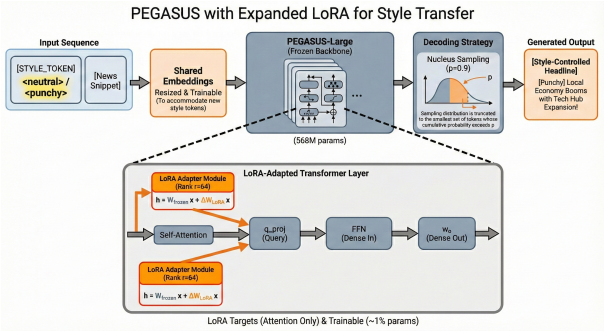

Fig. 1. LoRA adapter placement for the Attention-Only (A-Only) configuration, applied to the attention projections (Q_proj and V_proj).

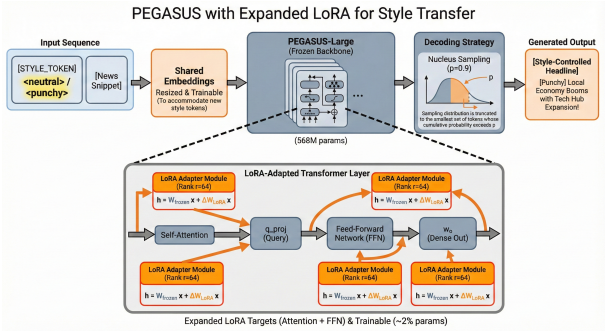

Fig. 2. LoRA adapter placement for the FFN-Expanded (FFN-E) configuration, adding adapters to the feed-forward layers (fc1 and fc2) in addition to the attention projections (Q_proj and V_proj).

The specific configurations tested are summarized in Table I. This range of designs allowed us to determine whether the observed failure mode, Stylistic Collapse, is dependent on adapter rank, module placement, or a more fundamental limitation of the PeFT approach itself. The trainable parameter ratio ranged from ∼0.5% (r=8, A-Only) to ∼8.0% (r=64, FFN-E) of the total model weights.

TABLE I
LoRA Configuration Variants Tested

Configuration Set	Rank (r)	Target Modules
Attention-Only (A-Only)	8, 16, 32, 64	`q_proj`, `v_proj`
FFN-Expanded (FFN-E)	8, 16, 32, 64	`q_proj`, `v_proj`, `k_proj`, `out_proj`, `fc1`, `fc2`

C. Training Setup

Fine-tuning was performed using the Hugging Face Trainer API with the following hyperparameters:

- Objective Function: Cross-entropy loss
- Optimizer: AdamW
- Learning Rate: 2×10^{-4}
- Batch Size: 8 (GPU memory-constrained)
- Epochs: 5
- Training Format: `[STYLE_TOKEN]` [SNIPPET] → [HEADLINE]

Because cross-entropy does not explicitly reward stylistic divergence, this setup isolates the question of whether supervised fine-tuning with style tokens and LoRA capacity alone can induce reliable style control.

1) Decoding Strategy: During inference, we employed nucleus sampling (top-p) with $p = 0.9$ and temperature $T = 0.8$ to allow controlled stochastic generation. Maximum output length was set to 32 tokens. Nucleus sampling was chosen to provide the model with flexibility to explore stylistic variations while maintaining coherent output. Notably, even with this sampling-based approach that permits diversity, the model exhibited high rates of identical outputs across different style prompts, demonstrating that stylistic collapse occurs at the representation level rather than being solely a deterministic decoding artifact.

D. Evaluation Procedures

Model performance was evaluated on the held-out test set for all eight LoRA configurations using three complementary dimensions:

1) **Content Fidelity:** ROUGE-1 scores were computed to measure lexical and structural overlap with reference headlines.
2) **Style Control and Output Divergence:** A proxy classifier trained on human-authored examples predicted whether outputs reflected neutral or punchy style. The Identical Output Rate (IOR) quantified how often the model produced the exact same summary across different style prompts—a key indicator of stylistic collapse.
3) **Factuality:** A RoBERTa-based MNLI model evaluated whether generated headlines were entailed by the source snippet.

IV. Experiments and Results

We evaluated the model's performance on the held-out test set across all eight LoRA configurations detailed in Table I. We report results along three dimensions: Stylistic Control (measured by identical output rate), Content Fidelity, and the preservation of Factuality. While increasing LoRA capacity improved content scores, the central finding remained consistent across all configurations: a high rate of output homogeneity regardless of the input style prompt.

A. Overall Performance and Stylistic Collapse Rate

The Identical Output Rate (IOR), defined as the percentage of test snippets producing byte-identical headlines for ¡neutral¿ versus ¡punchy¿ prompts, serves as the primary diagnostic for stylistic collapse. Table II shows IOR remained high across all eight configurations, ranging from 62.50% (A-Only r=16) to 79.69% (A-Only r=64). Crucially, increasing LoRA rank and expanding to FFN layers did not consistently reduce collapse; the highest-capacity models (A-Only r=64: 79.69%, FFN-E r=32: 78.13%) exhibited among the highest rates.

TABLE II
Stylistic Collapse Dynamics: Identical output rates remain high and erratic across all LoRA capacity settings, indicating an inability to reliably control style.

Config Set	Rank	Trainable Params (%)	Identical Output Rate (%)	Punchy Acc (%)
A-Only	8	~ 0.3	65.63	26.56
A-Only	16	~ 0.5	62.50	28.13
A-Only	32	~ 1.0	64.06	28.13
A-Only	64	~ 2.1	79.69	28.13
FFN-E	8	~ 1.0	68.75	29.69
FFN-E	16	~ 2.0	70.31	25.00
FFN-E	32	~ 3.8	78.13	29.69
FFN-E	64	~ 7.5	75.00	31.25

Content fidelity, measured by ROUGE-1, generally improved with increased capacity. FFN-Expanded models achieved ROUGE-1 scores around 0.52 for neutral headlines (Table IV), comparable to state-of-the-art fine-tuned PEGASUS performance, demonstrating successful summarization adaptation despite stylistic failure.

1) Qualitative Examples: Table III presents representative examples illustrating stylistic collapse. In Example 1, the model produces identical outputs for both style prompts, demonstrating complete stylistic collapse despite clear differences in the human references. Example 2 shows a rare case of partial divergence, where the model introduces minor lexical variation but fails to capture the dramatic tone shift present in the punchy reference. Both examples were taken from the Rank 64, FFN-E LoRA configuration.

B. Content Fidelity and Factuality

Content fidelity scores increased with LoRA capacity, as shown in Table IV. The FFN-Expanded models (particularly r=16,32,64) achieved the highest ROUGE-1 scores for the neutral style, peaking at 0.527 for the FFN-E r=16 model.

Factuality, measured via NLI entailment, averaged above 0.85 across all configurations (Table IV), with no significant drop in FFN-Expanded versus A-Only models.

V. Discussion

The most significant finding is the persistent stylistic collapse across all LoRA configurations. Despite varying ranks and target modules, IOR remained consistently high (62.50–79.69%), with some high-capacity models showing the *worst* collapse. This pattern indicates that the base model's pretraining and strong extractive bias substantially dominate the parameter-efficient style signal. Even when adapters cover both attention and FFN layers with up to 8% trainable parameters (FFN-E r=64), the model largely ignores style tokens and converges to a single high-probability headline per snippet.

The observed collapse suggests that, in our PEGASUS-Large experiments on this dataset, LoRA-based fine-tuning may be insufficient for stylistic headline generation when stylistic differences are subtle and datasets are small. The high IOR indicates the model defaults to extractive-style summaries irrespective of prompts. Combined with high factuality (entailment > 0.85) and strong ROUGE scores (neutral $\sim$0.52), this suggests the optimization primarily favors content-preserving solutions, treating style as a weak secondary objective that can be safely ignored (Fig. 3).

Dataset ambiguity likely amplified the collapse. The proxy style classifier achieved only modest accuracy on human references (approximately 56%), indicating that neutral versus punchy distinctions were subtle and sometimes overlapping. Under such conditions, cross-entropy training may provide weak or contradictory gradients for style tokens, further incentivizing the model to prioritize content fidelity over stylistic differentiation. This suggests a limitation in this setting: LoRA alone may not reliably enforce stylistic divergence without consistently separable style signals in both the dataset and objective function.

TABLE III
QUALITATIVE EXAMPLES SHOWING STYLISTIC COLLAPSE. EXAMPLE 1 DEMONSTRATES COMPLETE COLLAPSE (IDENTICAL OUTPUTS), WHILE EXAMPLE 2 SHOWS MINIMAL DIVERGENCE THAT FAILS TO CAPTURE THE INTENDED STYLISTIC CONTRAST.

Component	Text
Example 1: Complete Collapse	
Input Snippet	EIGHT people from the Department of Public Works and Highways (DPWH) linked to an anomalous Oriental Mindoro flood control project are now under government custody, with eight — including former Ako Bicol congressman Zaldy Co — still at large and believed to be abroad.
Neutral Reference	Eight DPWH Officials in Oriental Mindoro Flood Control Case Now in Custody
Punchy Reference	Mindoro Flood Graft Suspects: 8 Nabbed, 8 Still on the Run
Model Output (Neutral)	Eight DPWH Officials in Custody Over Oriental Mindoro Flood Control Project
Model Output (Punchy)	**Eight DPWH Officials in Custody Over Oriental Mindoro Flood Control Project**
Example 2: Minimal Divergence	
Input Snippet	The 2023 FIBA World Cup saw record-breaking attendance, especially in games featuring the Philippines.
Neutral Reference	FIBA World Cup games in PH see high attendance
Punchy Reference	PH fans flood the arenas — World Cup crowds set records
Model Output (Neutral)	FIBA World Cup 2023 breaks attendance records
Model Output (Punchy)	Record attendance at FIBA World Cup 2023

TABLE IV
CONTENT FIDELITY (ROUGE-1) AND FACTUALITY (ENTAILMENT) ACROSS ALL CONFIGURATIONS. HIGH SCORES IN THESE METRICS DEMONSTRATE SUCCESSFUL SUMMARIZATION ADAPTATION DESPITE STYLISTIC FAILURE.

Config Set	Rank	ROUGE-1 (Neutral)	ROUGE-1 (Punchy)	Entailment (Neutral)	Entailment (Punchy)
A-Only	8	0.459	0.319	0.904	0.922
A-Only	16	0.464	0.365	0.934	0.936
A-Only	32	0.461	0.361	0.953	0.909
A-Only	64	0.497	0.375	0.899	0.915
FFN-E	8	0.492	0.371	0.953	0.938
FFN-E	16	0.527	0.387	0.884	0.922
FFN-E	32	0.519	0.402	0.877	0.858
FFN-E	64	0.522	0.402	0.877	0.880

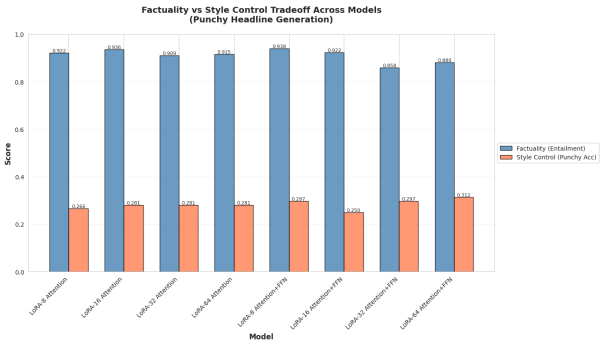

Fig. 3. A diagnostic comparison showing high factual consistency (Entailment) alongside poor stylistic control (Punchy Accuracy), indicating the dominance of the base model's summarization bias over the parameter-efficient style adaptation.

VI. CONCLUSION AND FUTURE WORK

This study shows that, in our PEGASUS-Large setting, parameter-efficient fine-tuning using LoRA may be insufficient for reliable style-controlled headline generation on PEGASUS-Large. Across eight configurations varying rank (r $\in \{8,16,32,64\}$) and target modules (Attention-Only vs. FFN-Expanded), the model exhibited severe stylistic collapse, with Identical Output Rate consistently exceeding 62%. Despite maintaining high factual consistency (entailment > 0.85) and competitive neutral-style ROUGE scores (0.52), the model ignored stylistic tokens and converged to near-identical outputs.

These findings confirm that increasing LoRA capacity—through higher rank or FFN expansion—cannot overcome the base model's dominant extractive bias when relying solely on cross-entropy loss. The failure persisted even with 8% trainable parameters, indicating a fundamental limitation of standard PEFT approaches for subtle style control.

Future work should first test whether similar collapse occurs in other architectures and tasks (e.g., BART/T5-style encoder–decoders, other generation objectives). Beyond that, alternatives that explicitly reward stylistic divergence: contrastive or divergence-based losses that penalize identical outputs across style prompts, reinforcement learning with style-specific reward functions, or multitask architectures that jointly optimize content preservation and stylistic separation. Additionally, investigating whether larger datasets with clearer stylistic boundaries can provide stronger training signals for PEFT methods represents a valuable direction.

ACKNOWLEDGMENT

The authors would like to thank the University of Asia and the Pacific, School of Sciences, Engineering and Technology (SSE) for its support. This work was done in partnership with the DOST-NAIRA project and was funded and monitored

by the Department of Science and Technology (DOST)–Philippine Council for Industry, Energy, and Emerging Technology Research and Development (PCIEERD) under Project No. 1213385.

REFERENCES

[1] Ficler, J., & Goldberg, Y. (2017). Controlling Linguistic Style Aspects in Neural Language Generation. *Proceedings of the Workshop on Stylistic Variation*.

[2] Jin, D., et al. (2022). Deep Learning for Text Style Transfer: A Survey. *Computational Linguistics*, 48(2).

[3] Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *International Conference on Machine Learning (ICML)*.

[4] Hu, E. J., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations (ICLR)*.

[5] Dettmers, T., et al. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. *Conference on Neural Information Processing Systems (NeurIPS)*.

[6] Lou, N., Song, H., Shang, W., Liu, X., Li, Z., Dong, Y., *et al.* (2023). Parameter-Efficient Fine-Tuning of Large-Scale Pre-trained Language Models: A Survey. *arXiv preprint arXiv:2303.15647*.

[7] Zhang, H., Song, H., Li, S., Zhou, M., & Song, D. (2023). A Survey of Controllable Text Generation Using Transformer-based Pre-trained Language Models. *ACM Computing Surveys*, 56(3).

[8] Jin, D., et al. (2021). Hooks in the Headline: Learning to Generate Headlines with Controlled Styles. *Annual Meeting of the Association for Computational Linguistics (ACL)*.

[9] Cao, S., & Wang, L. (2021). Inference-Time Style Control for Summarization. *arXiv preprint arXiv:2104.09503*.