Guide-to-Explain for Controllable Summarization

Sangwon Ryu¹, Heejin Do¹, Daehee Kim¹, Yunsu Kim³, Gary Geunbae Lee¹², Jungseul Ok¹²

¹Graduate School of Artificial Intelligence, POSTECH, South Korea

²Department of Computer Science and Engineering, POSTECH, South Korea

³aiXplain Inc., Los Gatos, CA, USA

{ryusangwon, heejindo, andrea0119, gblee, jungseul}@postech.ac.kr,
yunsu.kim@aixplain.com

Abstract

Recently, large language models (LLMs) have demonstrated remarkable performance in abstractive summarization tasks. However, controllable summarization with LLMs remains underexplored, limiting their ability to generate summaries that align with specific user preferences. In this paper, we first investigate the capability of LLMs to control diverse attributes, revealing that they encounter greater challenges with numerical attributes, such as length and extractiveness, compared to linguistic attributes. To address this challenge, we propose a guideto-explain framework (GTE) for controllable summarization. Our GTE framework enables the model to identify misaligned attributes in the initial draft and guides it in explaining errors in the previous output. Based on this reflection, the model generates a well-adjusted summary. As a result, by allowing the model to reflect on its misalignment, we generate summaries that satisfy the desired attributes in surprisingly fewer iterations than other iterative methods solely using LLMs.

1 Introduction

Large language models (LLMs) have demonstrated superior performance in abstractive summarization, outperforming traditional encoder-decoder models by generating more contextually appropriate and natural summaries (Goyal et al., 2023; Zhang et al., 2024; Pu et al., 2023; Ryu et al., 2024b). In addition, recent studies aimed to generate higher-quality summaries by leveraging the self-correction capabilities of LLMs (Zhang et al., 2023a; Sun et al., 2024). However, given individuals' diverse preferences for summary styles, it is essential to generate summaries that adjust personal needs (Zhang et al., 2023b). For instance, some users may prefer concise summaries or retain exact phrases from the original text.

Therefore, controllable summarization has recently garnered attention (Zhong et al., 2021; Xu

et al., 2023; Zhang et al., 2023b). Previous research employed encoder-decoder models to control attributes (Mao et al., 2022; Zhang et al., 2022; Vig et al., 2022; Pagnoni et al., 2023; Wang et al., 2023; Urlana et al., 2024). Although LLMs excel in generating high-quality summaries, they still face challenges in controlling attributes (Yuan et al., 2024; Tang et al., 2023), and their controllability has been underexplored (Liu et al., 2024).

Thus, we analyze LLMs' ability to control various attributes in summarization and refine the measurements to more accurately assess these attributes. We reveal that while LLMs excel at controlling linguistic attributes such as topic and speaker, they severely struggle with numerical attributes such as extractiveness and length. To address this challenge, we propose a guide-to-explain (GTE), which enables precise attribute control solely through LLMs without relying on external modules or training. We first design an attributeidentification step to calculate misaligned attributes in LLM-generated summaries, subsequently guiding the model to explain the sources of its errors. By self-reflecting its own errors, the model can adequately adjust attributes in subsequent iterations. We introduce the self-refine strategy, primarily used in reasoning tasks with LLMs (Weng et al., 2023; Madaan et al., 2023; Dhuliawala et al., 2024; Gou et al., 2024), to controllable summarization.

Additionally, we evaluate GTE on mixed attribute control datasets, $MACSum_{Doc}$ and $MACSum_{Dial}$ (Zhang et al., 2023b). GTE successfully controls each attribute with minimal iterations solely using LLMs, outperforming other iteration methods. We also demonstrate the high quality of the controlled summaries via multiple evaluation metrics. In addition, we analyzed whether LLMs can control multiple attributes simultaneously. We found out that LLMs struggle with jointly controlling correlated numerical attributes. Our contributions are as follows:

- We revisit the measurement of various attributes and analyze LLMs' controllability.
- We propose a guide-to-explain (GTE), which guides the model to explain its misalignments.
- GTE effectively adjusts misaligned attributes only in a few iterations.

2 Evaluating controllability of LLMs

2.1 Controllable attributes

We investigate the controllability of LLMs for four attributes: *extractiveness*, *length*, *topic*, and *speaker*. *Extractiveness* evaluates how much of the summary's content directly overlaps with the original text. A highly extractive summary is required when users need to retain the original context, such as in academic papers; however, paraphrasing is applied to tailor the summary in general cases. The *length* counts the ratio between the main text and the summary. The preferred summary length varies depending on the information density of the text and user preferences. For *topics* or *speakers*, users may prefer summaries focused on a specific topic or speakers from a long document or dialogue.

Attribute measurement Previous methods have not effectively accounted for attribute-focused aspects. In the case of extractiveness, it is straightforward to determine how much of the summary's content directly overlaps with the original text. However, for *length*, prompts suggested by earlier works specify a fixed number of sentences, e.g., "3 sentences," but this approach fails to account for variations in sentence length and does not accurately reflect the summary's actual length (Goyal et al., 2023; Liu et al., 2024; Yuan et al., 2024). Thus, we calculate the summary length as a ratio relative to the main text. For topics, Zhang et al. (2023b) calculated the frequency of topic-related words in the summary. However, even if topic words do not explicitly appear, the summary can still reflect the core context of the topic, especially for LLMgenerated summaries, which tend to paraphrase the content. Therefore, we compute the embedding similarity \mathcal{B} between the topic word and each word in the summary $s: \frac{1}{n} \sum_{i \in s} \mathcal{B}(topic, word_i),$ where n is the number of words in the summary. If multiple topics k are present, we use the average embedding similarity across all topics: $\frac{1}{k}\sum_{j\in k}\frac{1}{n}\sum_{i\in s}\mathcal{B}(topic_k,word_i)$. For speakers, Zhang et al. (2023b) calculate the frequency of the

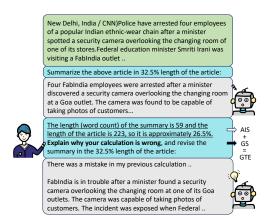


Figure 1: Overview of guide-to-explain system (GTE).

speaker's spoken words appearing in the summary. In contrast, we extract speaker utterances and construct speaker set \mathcal{S} and utilize BERTScore (Zhang et al., 2020) to compute the embedding similarity between the summary p and \mathcal{S} : BERTScore (p, \mathcal{S}) .

Label reinterpretation We use the publicly available MACSum dataset (Zhang et al., 2023b) for controllable summarization. However, existing labels for numerical attributes are ambiguous as criteria degrees are not specified (e.g., high for extractiveness and short for length). To provide detailed criteria, we reinterpret the labels based on the attribute distributions in each training set. For extractiveness, we set labels as {normal: 85%, high: 90%, fully: 100%}. Unlike previous methods, we define the summary length as a ratio of the original text rather than a fixed value. Since the existing labels do not distinctly differentiate these ratios, we set the lengths to {short: 7.5%, normal: 15%, long: 32.5%}, providing clearer distinctions with evaluating a broader range of controllability. Labeling details can be found in Appendix A.

2.2 Controllablility assessment

We evaluate the ability of LLMs to adjust their outputs based on specified attributes (Table 1, 2). Our assessment includes two evaluations: (1) the failure rate upon reaching the predefined maximum iterations without achieving the desired modifications and (2) the average iterations required to adjust an attribute, calculated only for successful cases. We denote the naive iteration approach, which simply adjusts attributes repeatedly, as Iter. Most LLMs effectively control linguistic attributes, such as *topic* and *speaker*. However, LLMs struggle with numerical attributes, including *extractiveness* and *length*. Both Llama-70B and GPT-40-Iter

Model	Extractiveness $(\downarrow / \downarrow)$				Length $(\downarrow / \downarrow)$				Topic(↓)	Speaker(↓)
	normal	high	full	avg	short	normal	long	avg	F-2(4)	Speaker(\psi)
Phi-3-medium-Iter	100.00% / 🖰	100.00% / 💍	100.00% / 🖰	100.00% / 🖰	100.00% / 🖰	100.00% / 🖰	100.00% / 🖰	100.00% / 🖰	38.08% / 0.22	-
Phi-3-medium-GTE	100.00% / 🖰	100.00% / 🖰	100.00% / 🖰	100.00% / 🖰	100.00% / 🖰	100.00% / 🖰	100.00% / 🖰	100.00% / 🖰	37.97% / 0.04	-
Llama3-8B-Iter	100.00% / 🖰	100.00% / 🖰	100.00% / 💍	100.00% / 🖰	100.00% / 🖰	100.00% / 🖰	100.00% / 🖰	100.00% / 🖰	57.14% / 0.12	-
Llama3-8B-GTE	12.63% / 3.52	11.63% / 2.53	0.00% / 1.46	11.70% / 3.26	26.40% / 3.08	10.92% / 2.26	13.18% / 3.85	14.99% / 2.80	25.56% / 0.91	-
Llama3-70B-Iter	54.82% / 8.44	37.21% / 7.47	2.70% / 3.78	49.91% / 8.05	18.40% / 6.58	54.61% / 10.42	67.44% / 12.00	49.36% / 8.24	0.00% / 0.24	-
Llama3-70B-AIS	26.55% / 6.57	18.60% / 7.81	0.00% / 1.86	24.14% / 6.52	4.80% / 5.42	2.73% / 3.81	10.85% / 4.84	5.12% / 4.39	0.00% / 0.10	-
Llama3-70B-GTE	0.21% / 3.28	0.00% / 2.83	0.00% / 1.50	0.18% / 3.22	0.00% / 1.10	0.00% / 1.61	2.32 % / 3.14	0.55% / 1.90	0.00% / 0.01	-
GPT-3.5-Iter	45.18% / 9.80	60.47% / 🖰	94.59% / 🖰	49.73% / 9.80	53.60% / 🖰	80.89% / 🖰	88.37% / 🖰	76.42% / 🖰	0.00% / 0.00	-
GPT-3.5-GTE	17.56% / 3.86	51.16% / 5.00	67.57% / 4.00	23.58% / 3.90	5.60% / 4.63	44.03% / 6.62	78.29% / 7.00	43.33% / 5.95	0.00% / 0.00	-
GPT-4o-Iter	34.69% / 6.77	55.81% / 🖰	78.38% / 3.00	39.31% / 6.63	72.00% / ひ	64.85% / 🖰	79.07% / 💍	69.84% / 🖰	0.38% / 0.02	-
GPT-4o-AIS	35.12% / 5.50	48.84% / 15.50	62.16% / 6.00	38.03% / 6.13	60.00% / 8.79	61.09% / 9.40	78.29% / 2.00	64.90% / 8.60	0.00% / 0.04	-
GPT-4o-GTE	0.00% / 2.76	0.00% / 4.70	0.00% / 2.03	0.00% / 2.87	0.00% / 1.20	0.00% / 1.21	0.00% / 1.96	0.00% / 1.42	0.00% / 0.02	-

Table 1: The results of controllability measured on the $MACSum_{Doc}$ dataset. The bold denotes the best performance. We set the maximum number of I terations to 20. If all the cases remained unadjusted in the initial draft and still not be resolved after reaching the maximum iterations, they were marked with \circlearrowright .

Model	Extractiveness $(\downarrow / \downarrow)$			Length (↓ / ↓)				Topic (\(\psi \)	Speaker (1)	
	normal	high	fully	avg	short	normal	long	avg	Topic (4)	
Llama3-70B-Iter	31.78% / 8.13	43.59% / 8.40	8.16% / 5.39	29.63% / 7.59	12.00% / 🖰	93.75% / 6.00	98.00% / 🖰	81.79% / 6.00	0.00% / 0.01	0.00% / 0.00
Llama3-70B-AIS	14.41% / 5.91	23.08% / 5.31	0.00% / 3.72	13.27% / 5.50	0.00% / 1.25	62.05% / 5.70	92.00% / 9.33	57.10% / 5.62	0.00% / 0.02	0.00% / 0.00
Llama3-70B-GTE	0.00% / 2.31	0.00% / 2.56	4.08% / 3.64	0.61% / 2.49	0.00% / 1.00	36.61% / 4.73	80.00% / 5.70	37.65% / 4.53	0.00% / 0.01	0.00% / 0.00
GPT-4o-Iter	79.24% / 4.36	82.05% 3.67	59.18% / 1.00	76.54% / 4.00	6.00% / 🖰	98.21% / 🖰	100.00% / 🖰	84.26% / 🖰	0.31% / 0.01	0.00% / 0.00
GPT-4o-AIS	84.75% / 4.00	87.18% 1.50	53.06% 5.10	80.25% / 4.32	2.00% / 4.50	96.43% / 🖰	100.00% / 🖰	82.41% / 4.50	0.00% / 0.01	0.00% / 0.00
GPT-4o-GTE	17.80% / 7.94	25.64% / 7.92	8.16% / 4.58	17.28% / 7.53	0.00% / 1.40	9.82% / 2.75	44.00% / 4.21	13.58% / 2.90	0.00% / 0.02	0.00% / 0.00

Table 2: The results of controllability measured on the MACSum $_{Dial}$ dataset.

show a very high failure rate with numerous iterations. In particular, LLMs struggle to adjust for normal *extractiveness* and long *length* attributes.

3 GTE

Therefore, we introduce a guide-to-explain (GTE) framework (Figure 1) to adjust the challenging numerical attributes. We provide an attribute identification step to adjust incorrectly generated responses and guide the LLMs to reflect by explaining the reasons behind these errors. Our approach allows the model to make appropriate corrections in subsequent iterations.

3.1 Attribute identification step

We first prompt the LLM to generate an initial draft s' that reflects the specified attributes i. If the LLM fails to control the instructed attributes accurately, we provide attribute identification step (AIS) to guide the model on how to adjust the attributes. LLM may have difficulty measuring attributes such as *extractiveness* or the *length* ratio relative to the original text. Thus, we provide a step-by-step approach to instruct the model on revising its summary based on the generated output.

3.2 Guidance step

After providing instructions AIS on revising the summary, we supply guidance step (GS) to the model to explain why it initially failed to adjust the attributes correctly. This process is similar to

how humans solve complex problems by reviewing their mistakes to produce more accurate responses in the future. Receiving [a; i, s'; AIS; GS] as inputs, where a denotes the article, the model first reflects on the reasons for the initial error before generating a revised summary. If the revised summary still fails to satisfy the attributes, GTE repeats until the model produces an attribute-compliant summary. As LLMs are known to struggle with number-related tasks (Thawani et al., 2021; Imani et al., 2023), our guidance to explain why their calculation is incorrect, followed by generating summaries assists in effectively controlling the numeric attributes. The used prompts are in Appendix B.

4 Experimental setup

We evaluate the controllability of various LLMs, including Phi-3 (Abdin et al., 2024), Llama3 (Dubey et al., 2024), and GPT series (Brown, 2020; Achiam et al., 2023). To analyze model performance by size, we use both the 8B and quantized 70B versions ¹ of Llama3, and GPT-3.5 and GPT-4.

We used two datasets, the $MACSum_{Doc}$ and the $MACSum_{Dial}$ datasets (Zhang et al., 2023b), which comprise committee meetings and news contents each. Both datasets are for mixed-attribute summarization that control multiple attributes simultaneously, but only $MACSum_{Dial}$ has speaker attribute. Since we evaluate LLM performance on

¹casperhansen/llama-3-70b-instruct-awq

Model	Coherence	Consistency	Fluency	Relevance	QuestEval
Iter (Ext)	0.820	0.800	0.859	0.696	0.523
AIS (Ext)	0.884	0.843	0.905	0.785	0.554
Iter (Len)	0.836	0.803	0.836	0.759	0.484
AIS (Len)	0.934	0.834	0.942	0.887	0.548
GTE (Ext)	0.941	0.873	0.937	0.880	0.590
GTE (Len)	0.937	0.840	0.944	0.901	0.553

Table 3: Quality of the controlled summaries.

individual attributes, we use attributes separately.

5 Results and Discussions

Main results We define the strategy solely providing attribute identification steps as AIS and denote our full guiding framework as GTE. As a result, our GTE demonstrates remarkably lower failure rates and fewer iterations when adjusting summaries across all attributes, including challenging numerical attributes in MACSum $_{Doc}$ (Table 1). Notably, while applying GTE to smaller models such as Phi-3 and Llama3-8B resulted in significant performance improvements, we observed that failures were almost nonexistent when applied to larger models such as Llama-70B or GPT-40.

LLMs encounter more difficulties with the MACSum_{Dial} dataset (Table 2). The dataset, which is derived from QMSum (Zhong et al., 2021), consists of lengthy and diverse content parliamentary and committee meetings, making it more challenging compared to the CNN-news-based MACSum_{Doc}. Notably, both the GPT-40-Iter and GPT-40-AIS failed to adjust for long length, whereas our GPT-40-GTE demonstrated a commendable success rate. Regarding extractive-ness, the Iter and AIS of GPT-40 exhibit relatively low iteration counts since the models mostly exceed the maximum iteration (①). While they fail nearly 80%, our GTE demonstrates a significantly lower failure rate at 17.28% with low iterations.

Quality of controlled summary Instead of ROUGE score(Lin, 2004), which does not adequately evaluate the quality of the summaries (Zhong et al., 2022; Scialom et al., 2021; Ryu et al., 2024a), we use UniEval (Zhong et al., 2022) and QuestEval (Scialom et al., 2021) to assess both the inherent quality of the summaries and their factual consistency with the source text. UniEval, a multi-dimensional evaluator with high human correlation, assesses dimensions such as *coherence*, *consistency*, *fluency*, and *relevance*, and QuestEval evaluates factuality via question answering. Table 3 shows that our method's summaries outperform all

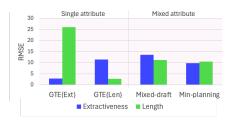


Figure 2: Performance in mixed-attribute.

UniEval dimensions and QuestEval scores, demonstrating effective attribute control while maintaining overall summary quality.

Mixed attributes controllability We observed that when generating summaries controlled for all attributes simultaneously, the model effectively handled linguistic attributes but faced challenges with numerical attributes. Notably, satisfying all attributes within the maximum number of iterations proved challenging for all methods, including GTE.

Sequential-planning Discovering the challenges in precisely controlling all attributes in parallel, we introduce a sequential-planning strategy, which gradually adjusts attributes from the ill-controlled with the initial draft using GTE. However, modifying one attribute often disrupted previously adjusted attributes due to correlations. For example, controlling *length* first would still lead to changes in *length* when adjusting *extractiveness*. Consequently, sequential adjustments result in a modest performance gap compared to the initial draft (Figure 2). We assess the attributes using the root mean squared error (RMSE) between the instructed values and those in the generated summaries.

6 Conclusion

In this work, we revisit the measurement of controllable summarization with various attributes. We evaluate the controllability of multiple attributes in summary generation with LLMs, finding that LLMs struggle to adjust numeric attributes compared to linguistic ones. To address this limitation, we propose a guide-to-explain (GTE) approach, where the model is guided to explain its misalignments and then grounded this explanation to produce better-controlled summaries in subsequent iterations. GTE enables LLMs to control challenging numerical attributes with lower failure and fewer iterations. Further, we validate the quality of the controlled summaries via a multi-dimensional evaluation, demonstrating the high-quality generation.

Limitation

We evaluated the controllability of various attributes in LLMs and introduced a novel guide-to-explain (GTE) framework to address challenges in numerical attributes. While GTE enhanced control over numerical attributes, it still struggled with highly correlated mixed numerical attributes. Additionally, sequential planning, which adjusts attributes in order of least alignment, also faced difficulties achieving precise control. Even after properly adjusting one attribute, modifying the correlated numerical attribute caused the previously adjusted attribute to change. We believe further research could explore more effective methods for addressing these challenges.

Ethics

We used publicly available MACSum datasets for our research, conducting experiments with Phi-3, Llama3 ², GPT-3.5, and GPT-40 from April to October 2024.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00217286) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH)).

References

- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv* preprint arXiv:2404.14219.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
- Alan Akbik, Bergmann, et al. 2019. FLAIR: An easy-touse framework for state-of-the-art NLP. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*.
- ²Meta Llama3 Community License, Copyright © Meta Platforms, Inc. All Rights Reserved. More details can be found at: Llama3 License

- Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3563–3578, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv* preprint arXiv:2407.21783.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.
- Tanya Goyal et al. 2023. News summarization and evaluation in the era of gpt-3. *Preprint*, arXiv:2209.12356.
- Shima Imani, Liang Du, and Harsh Shrivastava. 2023. MathPrompter: Mathematical reasoning using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 37–42, Toronto, Canada. 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.
- Yixin Liu, Alexander Fabbri, Jiawen Chen, Yilun Zhao, Simeng Han, Shafiq Joty, Pengfei Liu, Dragomir Radev, Chien-Sheng Wu, and Arman Cohan. 2024. Benchmarking generation and evaluation capabilities of large language models for instruction controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4481–4501, Mexico City, Mexico. Association for Computational Linguistics.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.
- Ziming Mao, Chen Henry Wu, Ansong Ni, Yusen Zhang, Rui Zhang, Tao Yu, Budhaditya Deb, Chenguang Zhu, Ahmed Awadallah, and Dragomir Radev. 2022. DYLE: Dynamic latent extraction for abstractive long-input summarization. In *Proceedings of the*

- 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1687–1698, Dublin, Ireland. Association for Computational Linguistics.
- Artidoro Pagnoni, Alex Fabbri, Wojciech Kryscinski, and Chien-Sheng Wu. 2023. Socratic pretraining: Question-driven pretraining for controllable summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12737–12755, Toronto, Canada. Association for Computational Linguistics.
- Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.
- Sangwon Ryu, Heejin Do, Yunsu Kim, Gary Lee, and Jungseul Ok. 2024a. Multi-dimensional optimization for text summarization via reinforcement learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5858–5871, Bangkok, Thailand. Association for Computational Linguistics.
- Sangwon Ryu, Heejin Do, Yunsu Kim, Gary Geunbae Lee, and Jungseul Ok. 2024b. Key-element-informed sllm tuning for document summarization. In *Interspeech* 2024, pages 1940–1944.
- Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Shichao Sun, Ruifeng Yuan, Ziqiang Cao, Wenjie Li, and Pengfei Liu. 2024. Prompt chaining or stepwise prompt? refinement in text summarization. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7551–7558, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Yuting Tang, Ratish Puduppully, Zhengyuan Liu, and Nancy Chen. 2023. In-context learning of large language models for controlled dialogue summarization: A holistic benchmark and empirical analysis. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 56–67, Singapore. Association for Computational Linguistics.
- Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. 2021. Representing numbers in NLP: a survey and a vision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–656, Online. Association for Computational Linguistics.
- Ashok Urlana, Pruthwik Mishra, Tathagato Roy, and Rahul Mishra. 2024. Controllable text summarization: Unraveling challenges, approaches, and

- prospects a survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1603–1623, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Jesse Vig, Alexander Fabbri, Wojciech Kryscinski, Chien-Sheng Wu, and Wenhao Liu. 2022. Exploring neural models for query-focused summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1455–1468, Seattle, United States. Association for Computational Linguistics.
- Bin Wang, Zhengyuan Liu, and Nancy Chen. 2023. Instructive dialogue summarization with query aggregations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7630–7653, Singapore. Association for Computational Linguistics.
- Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2550–2575, Singapore. Association for Computational Linguistics.
- Ruochen Xu, Song Wang, Yang Liu, Shuohang Wang, Yichong Xu, Dan Iter, Pengcheng He, Chenguang Zhu, and Michael Zeng. 2023. LMGQS: A large-scale dataset for query-focused summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14764–14776, Singapore. Association for Computational Linguistics.
- Weizhe Yuan, Ilia Kulikov, Ping Yu, Kyunghyun Cho, Sainbayar Sukhbaatar, Jason Weston, and Jing Xu. 2024. Following length constraints in instructions. *Preprint*, arXiv:2406.17744.
- Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023a. SummIt: Iterative text summarization via ChatGPT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10644–10657, Singapore. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of the International Conference on Learning Representations*.
- Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.
- Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir Radev, Chenguang Zhu, Michael Zeng, and Rui Zhang. 2023b. MACSum: Controllable summarization with mixed attributes. *Transactions of the Association for Computational Linguistics*, 11:787–803.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. Summⁿ: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multidimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

A Attribute details

A.1 Attribute analysis

We analysis the data distributions in the $MACSum_{Doc}$ and $MACSum_{Dial}$ (Zhang et al., 2023b) datasets. The attributes used in the original datasets are described as follows:

- Extractiveness: Controls how much of the summary is directly extracted from the source text. It is evaluated using the average of ROUGE-2 and ROUGE-3 precision scores.
- *Length*: The number of words in the summary. It is evaluated based on token length.
- *Topic*: Control the summary to align with the given topic, and multiple topics can be present. The evaluation is based on the proportion of topic words appearing in the summary.
- *Speaker*: Control the summary to focus on the speech of a specific speaker. It is evaluated by the ratio of the speaker's words included in the summary.
- Specificity: Controls the level of detail or descriptive content in the summary. The evaluation formula is: $Specificity = (0.1 \times 10^{-5})$

 $verbs + 0.2 \times tokens + 0.3 \times nouns + 0.4 \times numerical\ tokens)/number\ of\ sentences.$

The detailed data distribution for both $MACSum_{Doc}$ and $MACSum_{Dial}$ are presented in Table 4. We observe that the labels in previous work do not account for attributes. In the case of topic, the specific topic keywords may not always appear explicitly, even if the summary is focused on the given topic. Similarly, for speaker, LLMs may paraphrase the speaker's words using different terms or expressions. Therefore, we reinterpreted these labels in Section 2.1 to account for these factors.

For the *extractiveness* and *length* attributes, the original labels were categorized as 'normal', 'high', 'fully' for *extractiveness* and 'short', 'normal', 'long' for *length*. However, since the labels were not numerically defined, we deemed the criteria unclear. Therefore, we redefined these labels for our use.

We calculate *extractiveness* based on the proportion of words in the source text that appeared exactly in the summary. For *length*, while the original dataset was labelled based on word count and compression rate, we only considered the ratio of the summary to the original text. Table 4 shows that the ratios were not strictly distinct, with some overlap between labels. Thus, we relabeled boundaries in Section 2.1. Since these label values are not exact for the reference summary, we adjusted attributes according to the specified new labels. Consequently, we did not use the reference summaries from the original dataset.

For *specificity*, we used the Flair ³ (Akbik et al., 2019) model for named entity recognition (NER) model to measure the number of entities in each summary. Since the number of entities varies with summary length, we calculated specificity as the ratio of total entities to total words in the summary. However, the annotation for specificity in the original dataset was inversely related to our calculations, so we did not use this attribute in our study.

A.2 Iteration threshold

We set attribute-specific thresholds and iteratively adjusted them until the criteria were satisfied. For *extractiveness* and *length*, iterations were considered successful if the values fell within a relabeled value ± 5 range. For the *topic* and *speaker*, we set the minimum embedding similarity score as

³https://github.com/flairNLP/flair

Attribute	Label]	MACSum	Doc	\mathbf{MACSum}_{Dial}			
Attinbute	Laber	Dist.	Relabel	# of summaries	Dist.	Relabel	# of summaries	
	short	0.7 - 15.0%	7.5%	1059	0.2 - 20.8%	7.5%	300	
Length	normal	0.5 - 48.6%	20.0%	2194	0.3 - 41.9%	20.0%	1693	
	long	1.5 - 39.8%	32.5%	1025	0.7 - 32.4%	32.5%	345	
	normal	35.7 - 100.0%	85.0%	3731	53.2 - 100.0%	85.0%	1661	
Extractiveness	high	55.0 - 100.0%	90.0%	287	63.0 - 100.0%	90.0%	340	
	fully	84.6 - 100.0%	100.0%	260	75.9 - 100.0%	100.0%	337	
Topic	-	74.8 - 88.8	74	2013	73.6 - 87.0	74	2317	
Speaker	-	-	-	-	75.6 - 92.0	75	1796	
G 'C'	normal	0.105%	-	3713	0.115%	-	1702	
Specificity	high	0.062%	-	565	0.110%	-	636	

Table 4: Data distribution of $MACSum_{Doc}$ and $MACSum_{Dial}$

the criterion for achieving a successful topic- or speaker-focused summary.

B Guide-to-explain (GTE) prompts

Article

(CNN)Easter is unique on the Christian calendar, a major point in the cycle of the religious year, and one that has always been able to resist the commercialization and culture warring that surrounds Christmas. That's in part because Easter is genuinely about how religious impulses, and patterns, can operate in ways that affect our lives. Nevertheless, I'm often surprised by how little people, even those supposedly within the Christian tradition, actually know about what is called Holy Week and its culmination on Easter Sunday. At a time when our culture is rolled by questions of identity and ethics (and tolerance) that have profound religious implications, it's worth pausing to explore this crucial holiday -- and the awareness of the human condition, in all its sadness and glory, that it engenders. After all, Holy Week calls mostly to those who incline their minds and hearts in its direction with seriousness of intent. Still, the fuss must puzzle those looking on, wondering what it all means. Why do Christians make so much of this springtime week, and make so much of Easter weekend? There is a phrase that many never come across, even among Christians: Easter Triduum. This refers to the three days of Easter that begin with Good Friday, proceed through Holy Saturday, and conclude with Easter Sunday. It's definitely a progression, although the word itself -- triduum -- can refer to any three days of prayer...

Initial prompt

summarize the above article in 32.5% length of the article:

Summary (16.9%)

Easter is a significant Christian holiday that often gets overlooked, especially when compared to Christmas. Nearly everyone within the Christian tradition knows the basics of Easter, but not many understand the full significance and rituals of Holy Week and the Easter Triduum, which are three days of prayer that begin on Good Friday and end on Easter Sunday. Holy Communion originated from the Last Supper, a Passover meal that Jesus shared with his disciples before being crucified. On Maundy Thursday, Jesus washed the disciples' feet to demonstrate humility...

GTE

The length (word count) of the summary is 157 and the length of the article is 931, so it is approximately 16.9%. Explain why your calculation is wrong, and revise the summary in the 32.5% length of the article:

Revised Summary (29.4%)

The calculation is wrong because the article length is 931 words, and the summary length is 157 words. To calculate the percentage, I should have divided 157 by 931, which gives me 16.9%. This means the summary is only 16.9% of the original article, not 32.5%.

To revise the summary to reach 32.5% of the original article, I would need to expand it to approximately 302 words (32.5% of 931). Here is a revised summary:

Easter is a significant Christian holiday that often gets overlooked, especially when compared to Christmas. Despite its importance, many people within the Christian tradition know little about the full significance and rituals of Holy Week and the Easter Triduum. This three-day period of prayer begins on Good Friday, commemorating Jesus' crucifixion, and concludes on Easter Sunday, celebrating his resurrection...

Figure 3: Length guide-to-explain (GTE).

Article

(CNN)Easter is unique on the Christian calendar, a major point in the cycle of the religious year, and one that has always been able to resist the commercialization and culture warring that surrounds Christmas. That's in part because Easter is genuinely about how religious impulses, and patterns, can operate in ways that affect our lives. Nevertheless, I'm often surprised by how little people, even those supposedly within the Christian tradition, actually know about what is called Holy Week and its culmination on Easter Sunday. At a time when our culture is roiled by questions of identity and ethics (and tolerance) that have profound religious implications, it's worth pausing to explore this crucial holiday -- and the awareness of the human condition, in all its sadness and glory, that it engenders. After all, Holy Week calls mostly to those who incline their minds and hearts in its direction with seriousness of intent. Still, the fuss must puzzle those looking on, wondering what it all means. Why do Christians make so much of this springtime week, and make so much of Easter weekend? There is a phrase that many never come across, even among Christians: Easter Triduum. This refers to the three days of Easter that begin with Good Friday, proceed through Holy Saturday, and conclude with Easter Sunday. It's definitely a progression, although the word itself -- triduum -- can refer to any three days of prayer...

Initial prompt

summarize the above article in 32.5% length of the article:

Summary (16.9%)

Easter is a significant Christian holiday that often gets overlooked, especially when compared to Christmas. Nearly everyone within the Christian tradition knows the basics of Easter, but not many understand the full significance and rituals of Holy Week and the Easter Triduum, which are three days of prayer that begin on Good Friday and end on Easter Sunday. Holy Communion originated from the Last Supper, a Passover meal that Jesus shared with his disciples before being crucified. On Maundy Thursday, Jesus washed the disciples' feet to demonstrate humility...

AIS

The length (word count) of the summary is 157 and the length of the article is 931, so it is approximately 16.9%. Summarize the above article in the 32.5% length of the article:

Revised Summary (23.4%)

Easter, a significant Christian holiday, remains untouched by commercialization and culture wars. Many people, including Christians, are unfamiliar with Holy Week and its culmination on Easter Sunday. The Easter Triduum, a three-day period, begins with Maundy Thursday, where Jesus shared the Last Supper with his disciples, demonstrating his humility by washing their feet. Good Friday, the day of the crucifixion, is a somber reminder of death, sacrifice, and fear. The Gospels provide varying accounts of Jesus' final words, including his quote from the 22nd Psalm. Holy Saturday, often overlooked, serves as a passage between the darkness of Good Friday and the hope of Easter Sunday. It's a day of anxiety and uncertainty, when the pain of the crucifixion still lingers, but hope begins to emerge. Easter Sunday, with its promise of transformation and resurrection, is a celebration of the human experience. The resurrection is not just a physical revival but a total transformation, symbolizing the possibility of change and renewal in our own lives...

Figure 4: Length problem-solving step (AIS).

Article

(CNN)Easter is unique on the Christian calendar, a major point in the cycle of the religious year, and one that has always been able to resist the commercialization and culture warring that surrounds Christmas. That's in part because Easter is genuinely about how religious impulses, and patterns, can operate in ways that affect our lives. Nevertheless, I'm often surprised by how little people, even those supposedly within the Christian tradition, actually know about what is called Holy Week and its culmination on Easter Sunday. At a time when our culture is roiled by questions of identity and ethics (and tolerance) that have profound religious implications, it's worth pausing to explore this crucial holiday -- and the awareness of the human condition, in all its sadness and glory, that it engenders. After all, Holy Week calls mostly to those who incline their minds and hearts in its direction with seriousness of intent. Still, the fuss must puzzle those looking on, wondering what it all means. Why do Christians make so much of this springtime week, and make so much of Easter weekend? There is a phrase that many never come across, even among Christians: Easter Triduum. This refers to the three days of Easter that begin with Good Friday, proceed through Holy Saturday, and conclude with Easter Sunday. It's definitely a progression, although the word itself -- triduum -- can refer to any three days of prayer...

Initial prompt

summarize the above article in 32.5% length of the article:

Summary (16.9%)

Easter is a significant Christian holiday that often gets overlooked, especially when compared to Christmas. Nearly everyone within the Christian tradition knows the basics of Easter, but not many understand the full significance and rituals of Holy Week and the Easter Triduum, which are three days of prayer that begin on Good Friday and end on Easter Sunday. Holy Communion originated from the Last Supper, a Passover meal that Jesus shared with his disciples before being crucified. On Maundy Thursday, Jesus washed the disciples' feet to demonstrate humility...

Iter

summarize the above article in 32.5% length of the article:

Revised Summary (9.1%)

Easter is a unique Christian event that resisted commercialization. Many Christians lack understanding of Holy Week and Easter. The Easter Triduum - Good Friday, Holy Saturday, and Easter Sunday - marks death, transition, and resurrection. This pattern reflects our daily struggles, suffering, and rebirth, making Easter a celebration of moving from darkness to light.

Figure 5: Length iteration (Iter).

Article

(CNN)We did it again, in another American city. We set Baltimore on fire this time. We brutalized black bodies. We turned a funeral into a riot. We let things get out of hand. We looted. We threw stones at policemen. We threw stones at citizens. We created camera-ready chaos, and we replayed the images. We created a culture of such deep distrust and disrespect that violence seemed the inevitable response. We let the violence flow. We let the violence stand for everything that's wrong with the things we already didn't like. By now you may be asking, "Who's we? "You may be saying with some irritation, "Don't lump me in with them. I didn't have anything to do with it." To which the only real answer can be: Stop kidding yourself. The word "we" is one of the great American words. We the People. Yes we can. We are family. I use "we" a lot when I talk about our country's achievements. I like to say we won the Second World War, we put a man on the moon, we invented the Internet, we gave the world jazz. Well, if I -- a son of immigrants whose family had nothing to do with any of those accomplishments -- if I get to claim those aspects of American history, then surely I have to claim the unsavory aspects too. "We" cuts both ways. We enslaved Africans. We cut Reconstruction short and made a mockery of equal citizenship. We supported Jim Crow, then redlined, subordinated, and ghettoized African-Americans...

Initial prompt

summarize the above article $\,$ using 85% of words directly from the original document:

Summary (100.0%)

We did it again, in another American city. We set Baltimore on fire this time. We brutalized black bodies. We turned a funeral into a riot. We let things get out of hand. We looted. We threw stones at policemen. We threw stones at citizens. We created camera-ready chaos, and we replayed the images. By now you may be asking, "Who's we?" You may be saying with some irritation, "Don't lump me in with them. I didn't have anything to do with it." To which the only real answer can be: Stop kidding yourself. The word "we" is one of the great American words. We the People...

GTE

This summary contains 100.0% of words directly from the original document. Explain why your calculation is wrong, and revise the summary so that 85% of the words are used directly as in the original document:

Revised Summary (86.8%)

I apologize for the mistake. Upon re-calculating, I realized that the original summary contained 79.4% of words directly from the original document. To revise the summary to meet the 85% threshold, I'll incorporate more words directly from the original document. Here's the revised summary:

We did it again, in another American city", perpetuating a cycle of violence and chaos. "We set Baltimore on fire this time", and in doing so, "we brutalized black podies". The author emphasizes the significance of "we", underscoring collective responsibility for both accomplishments and injustices. "We the People" mplies a shared history, encompassing achievements like "we won the Second World War", "we put a man on the moon", "we invented the Internet", and "we gave the world jazz", as well as darker aspects like "we enslaved Africans", "we cut Reconstruction short and made a mockery of equal citizenship", "we supported Jim Crow, then redlined, subordinated, and ghettoized African-Americans", and "we failed"...

Figure 6: Extractiveness guide-to-explain (GTE).

Article

(CNN)We did it again, in another American city. We set Baltimore on fire this time. We brutalized black bodies. We turned a funeral into a riot. We let things get out of hand. We looted. We threw stones at policemen. We threw stones at citizens. We created camera-ready chaos, and we replayed the images. We created a culture of such deep distrust and disrespect that violence seemed the inevitable response. We let the violence flow. We let the violence stand for everything that's wrong with the things we already didn't like. By now you may be asking, "Who's we? "You may be saying with some irritation, "Don't lump me in with them. I didn't have anything to do with it." To which the only real answer can be: Stop kidding yourself. The word "we" is one of the great American words. We the People. Yes we can. We are family. I use "we" a lot when I talk about our country's achievements. I like to say we won the Second World War, we put a man on the moon, we invented the Internet, we gave the world jazz. Well, if I -- a son of immigrants whose family had nothing to do with any of those accomplishments -- if I get to claim those aspects of American history, then surely I have to claim the unsavory aspects too. "We" cuts both ways. We enslaved Africans. We cut Reconstruction short and made a mockery of equal citizenship. We supported Jim Crow, then redlined, subordinated, and ghettoized African-Americans...

Initial prompt

summarize the above article using 85% of words directly from the original document:

Summary (100.0%)

We did it again, in another American city. We set Baltimore on fire this time. We brutalized black bodies. We turned a funeral into a riot. We let things get out of hand. We looted. We threw stones at policemen. We threw stones at citizens. We created camera-ready chaos, and we replayed the images. By now you may be asking, "Who's we?" You may be saying with some irritation, "Don't lump me in with them. I didn't have anything to do with it." To which the only real answer can be: Stop kidding yourself. The word "we" is one of the great American words. We the People...

AIS

This summary contains 100.0% of words directly from the original document. Summarize the above article using 85% of the words directly from the original document:

Revised Summary (78.0%)

Here's a rewritten summary that meets the 85% requirement:

We allowed Baltimore to burn, brutalizing black bodies and turning a funeral into a riot. We created chaos, replaying the images and fostering a culture of distrust and disrespect. When asked "Who's we?", some may object, saying "Don't lump me in with them." But the truth is, we as a nation are responsible. The word "we" is synonymous with American identity. If we claim America's achievements, we must also claim its dark past, including slavery and racial inequity. We are the authors of Baltimore's story, and it's our responsibility to address the problem. We can't just point to individual faults or structural flaws; we must acknowledge both. It's time to put aside our differences, acknowledge the truth, and work together to create opportunity and change.

Figure 7: Extractiveness problem-solving step (AIS).

Article

(CNN)We did it again, in another American city. We set Baltimore on fire this time. We brutalized black bodies. We turned a funeral into a riot. We let things get out of hand. We looted. We threw stones at policemen. We threw stones at citizens. We created camera-ready chaos, and we replayed the images. We created a culture of such deep distrust and disrespect that violence seemed the inevitable response. We let the violence flow. We let the violence stand for everything that's wrong with the things we already didn't like. By now you may be asking, "Who's we? "You may be saying with some irritation, "Don't lump me in with them. I didn't have anything to do with it." To which the only real answer can be: Stop kidding yourself. The word "we" is one of the great American words. We the People. Yes we can. We are family. I use "we" a lot when I talk about our country's achievements. I like to say we won the Second World War, we put a man on the moon, we invented the Internet, we gave the world jazz. Well, if I -- a son of immigrants whose family had nothing to do with any of those accomplishments -- if I get to claim those aspects of American history, then surely I have to claim the unsavory aspects too. "We" cuts both ways. We enslaved Africans. We cut Reconstruction short and made a mockery of equal citizenship. We supported Jim Crow, then redlined, subordinated, and ghettoized African-Americans....

Initial prompt

summarize the above article using 85% of words directly from the original document:

Summary (100.0%)

We did it again, in another American city. We set Baltimore on fire this time. We brutalized black bodies. We turned a funeral into a riot. We let things get out of hand. We looted. We threw stones at policemen. We threw stones at citizens. We created camera-ready chaos, and we replayed the images. By now you may be asking, "Who's we?" You may be saying with some irritation, "Don't lump me in with them. I didn't have anything to do with it." To which the only real answer can be: Stop kidding yourself. The word "we" is one of the great American words. We the People...

Iter

Summarize the above article using 85% of the words directly from the original document:

Revised Summary (100.0%)

We set Baltimore on fire this time. We brutalized black bodies. We turned a funeral into a riot. We let things get out of hand. We looted. We threw stones at policemen. We threw stones at citizens. The word "we" is one of the great American words. We the People. I use "we" a lot when I talk about our country's achievements. Well, if I get to claim those aspects of American history, then surely I have to claim the unsavory aspects too. "We" cuts both ways. We enslaved Africans. We cut Reconstruction short and made a mockery of equal citizenship. We supported Jim Crow, then redlined, subordinated, and ghettoized African-Americans. We cut blacks out of the New Deal. We created a polity in which racial inequity and economic inequality magnify each other unrelentingly. We tried to put a lid on it with heavy policing and a War on Drugs. We failed...

Figure 8: Extractiveness iteration (Iter).