ChatGPT Summarization: A Deep Dive into In-Context Learning Efficacy

Anonymous ACL submission

Abstract

Large language models (LLMs), such as ChatGPT, have risen to prominence in text summarization tasks, primarily due to the advent of in-context learning. This paper delves into how in-context learning steers the outputs of LLMs based on different data demonstration configurations. Our pivotal findings reveal that ChatGPT's adaptability to target summarization tasks is enhanced when provided with paired text and summaries, compared to when provided in isolation. Furthermore, the structured presentation of these pairs proves more influential than their precise content alignment. However, there are observable limitations: increasing the number of demonstrations yields diminishing returns, and the improvement of adaptability declines when tasked with more intricate news texts as opposed to simpler dialogues. This study comprehensively explains in-context learning's nuances in text summarization, highlighting its merits and demerits for future researchers.

1 Introduction

The burgeoning role of LLMs in text summarization underscores their growing significance in natural language processing. Pivotal to this surge is the advent of in-context learning (Brown et al., 2020), a technique that allows LLMs to adapt their outputs based on a handful of example demonstrations, thus obviating the need for extensive fine-tuning on specialized datasets. Recent studies (Zhang et al., 2023b; Yang et al., 2023; Zhang et al., 2023a; Wang et al., 2023) have highlighted its widespread use in text summarization. However, despite its success, a comprehensive exploration is still required to discern its influence on the generated summaries.

In-context learning involves crafting a prompt that directs the LLM toward a specific target task. The prompt initiates with an instruction defining the intended task and is followed by examples that guide the LLM in recognizing the task's structure. Once the foundation is set, the target input is incorporated into the prompt and fed into the LLM. Finally, the LLM endeavors to generate responses that address the target input and resonate with both the instructions and the provided examples. Refer to Table 1 for a showcase of prompts created for the summarization task with or without text-summary examples

In this study, we investigate the effects of in-context learning on the text summarization adaptability of LLMs, using ChatGPT as our primary model. We introduce four unique prompts, varying based on the inclusion or omission of text and summary, to determine their impact on ChatGPT's summarization performance. In a supplementary experiment, we manipulated text-summary alignments through text shuffling and replacement to explore the influence of content consistency on in-context learning. Our key findings include:

- Providing both texts and summaries to ChatGPT leads to better adaptation to summarization tasks than using either alone.
- The structured presentation of texts and summaries has a more pronounced impact than their exact content alignment.
- Increasing text-summary demonstrations show diminishing returns, highlighting the limitations of in-context learning.
- The efficacy of in-context learning decreases when summarizing complex text, as demonstrated by its greater effectiveness for dialogue than news summarization.

Our findings contribute to the broader understanding of in-context learning's role in text summarization, offering valuable insights for researchers and practitioners.

2 Related Work

The rise of LLMs has revolutionized the domain of natural language processing, particularly in the realm of text summarization. Historically, text summarization strategies leaned heavily on pre-trained models such as BART (Lewis et al., 2020), T5 (Raffel et al., 2023), and PEGASUS (Zhang et al., 2020). These models demanded resource-intensive fine-tuning to achieve optimal summary alignment with reference texts.

However, the advent of LLMs, including GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022), and LLaMA (Touvron et al., 2023), has introduced a paradigm shift. Instead of traditional fine-tuning, these models capitalize on their inherent capabilities, achieving remarkable results in zero-shot settings using well-crafted prompts with a few example demonstrations. This innovative approach, termed "In-context Learning" (Brown et al., 2020), has garnered significant attention recently.

Several studies have delved into the intricacies of in-context learning across diverse tasks. For instance, Xie et al. (2022) provides a theoretical perspective, suggesting that in-context learning can be conceptualized as Bayesian inference. From a practical standpoint, the work of Liu et al. (2022) emphasizes the pivotal role of example selection, illustrating that performance can be significantly enhanced when examples resonate closely with the target input. Furthermore, the research by Min et al. (2022a) highlights the importance of the structural integrity of examples, indicating that even flawed examples can bolster accuracy if presented structurally. On the other hand, Wei et al. (2023) delves into the nuances of model size, revealing that the effects of errors in examples can vary based on the dimensions of the model.

While in-context learning has seen widespread application in text summarization, notably in the news (Zhang et al., 2023b) and medical domains (Yang et al., 2023), as well as in extractive summarization (Zhang et al., 2023a), its specific impact on text summarization within LLMs requires deeper exploration. To the best of our knowledge, our study is the first to delve into how in-context learning influences LLMs in text summarization, aiming to bridge the existing research gap and offer crucial insights for subsequent investigations in the field.

3 Experimental Setup

3.1 Prompt Design

Prompt Name	**Prompt Structure**
No-Demo	Please provide a summary of the following text: {target_text} Summary:
Text-Summary	For reference on the desired summary style, here are separate texts and summaries: Texts: Example 1: {text_1} ... Summaries: Example 1: {summary_1} ... Given the reference examples, please provide a summary of the following text: {target_text} Summary:
Text-Only	For reference on text style: Example 1: {text_1} ... Given the reference examples, please provide a summary of the following text: {target_text} Summary:
Summary-Only	For reference on the desired summary style: Example 1: {summary_1} ... Given the reference examples, please provide a summary of the following text: {target_text} Summary:

Table 1: Detailed Prompt Structures. {target_text} is the target to be summarized. {text_*} and {summary_*} are the example texts and summaries.

Our experiment differentiates itself through specialized prompt designs when instructing ChatGPT. These designs include:

- *No-Demonstration (No-Demo)*, a direct prompt asking the model to summarize a provided text;
- *Text-Summary*, where a list of texts and a list of summaries are presented;
- *Text-Only*, offering only the example texts to hint at content style;
- *Summary-Only*, inverting the *Text-Only* by showcasing just the summaries, emphasizing the desired summary writing style.

Please refer to Table 1 for a clear representation of each prompt.

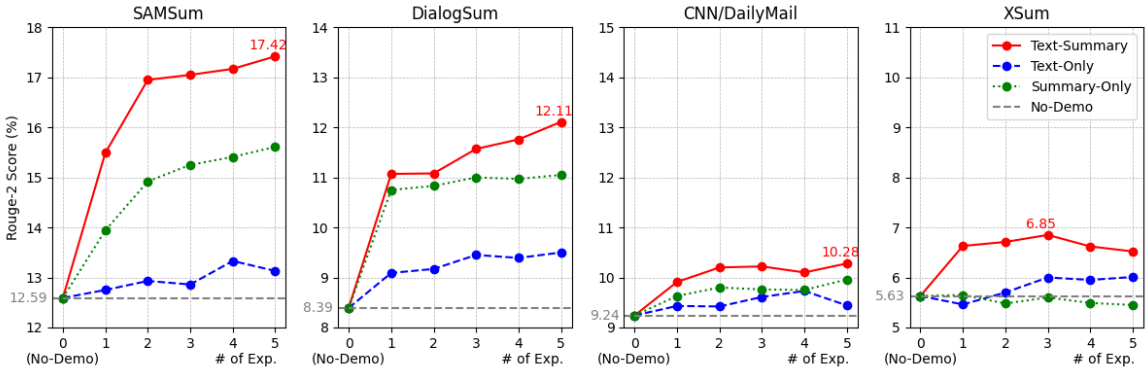

Figure 1: ROUGE-2 scores for the proposed prompts across four datasets, using one to five in-context examples. For each experiment, three separate trials are conducted using different sets of example data demonstrations. The results of these trials are then averaged. Note that *Text-Summary* prompt consistently achieves the highest scores. The value of Rouge-1, Rouge-2, and Rouge-L scores can be found in Table 4 in Appendix.

Dataset	Domain	#samples	Text #words	Summ. #words
SAMSum	Dialogue	819	94.6	20.0
DialogSum	Dialogue	1500	133.9	18.7
CNN/DM	News	1000	619.0	32.1
XSum	News	1000	388.7	21.2

Table 2: Overview of the datasets used for dialogue and news summarization, detailing the domain, sample sizes, and average word counts for the text and summary.

3.2 Datasets

We utilized four benchmark datasets: SAMSum (Gliwa et al., 2019) and DialogSum (Chen et al., 2021) for dialogue summarization, and CNN/DailyMail (Hermann et al., 2015) and XSum (Narayan et al., 2018) for news summarization. We fully utilized the available samples for testing, with 819 from SAMSum and 1500 from DialogSum. Meanwhile, we chose the first 1000 entries from XSum and CNN/DailyMail for our experiments. Table 2 provides an overview of the datasets.

3.3 ChatGPT

We used the *gpt-3.5-turbo-16k* version of the ChatGPT API from OpenAI for our experimental implementation. This version was chosen due to its capacity to handle up to 16000 tokens, making it suitable for the XSum and CNN/DailyMail datasets, which contain longer articles. Throughout our experiments, we adhered to the default parameter settings provided by ChatGPT to maintain consistency and to ensure that any observed behavior was attributable to the in-context learning and not to any custom configurations.

4 Experiment Results

4.1 Efficacy of Data Demonstrations

We initiated our empirical analysis by directing ChatGPT to perform a series of tests using the four prompts introduced in section 3.1 across the datasets. As depicted in Figure 1, the summarization performance, as measured by the Rouge-2 score, varied significantly based on the prompt format. *Text-Summary* demonstrated superior performance compared to *No-Demo*, highlighting the benefits of in-context learning to improve ChatGPT's adaptability to the target summarization task. While the *Summary-Only* showcased an improvement by adapting to reference summary styles, it did not match the efficacy of prompts integrating text and summary. On the other hand, the *Text-Only* presented only slight advancements over the *No-Demo*, emphasizing the importance of the text-summary relationship for optimal performance.

Upon further analysis of the datasets, we observed non-uniform improvements in summarization performance. The dialogue datasets, SAMSum and DialogSum, which are generally less intricate, experienced larger gains of 4.83% and 3.72%, respectively. In contrast, the more complex news datasets, CNN/DailyMail and XSum, displayed restrained advancements of 1.04% and 1.22%, respectively. This differential performance across datasets hints at the potential influences of text complexity on in-context learning efficacy.

Additionally, our experiments revealed a trend of diminishing returns from in-context learning: while 1-2 examples led to remarkable improvement, adding more examples resulted in a perfor-

Dataset	No-Demo	Original	Shuffled-Text	Replaced-Text
SAMSum	12.59	17.15	17.45	15.54
DialogSum	8.39	11.63	11.44	11.42
CNN/DailyMail	9.24	10.20	10.34	10.14
XSum	5.63	6.67	6.81	6.26

Table 3: Comparison of Rouge-2 scores for the *No-Demo* baseline and perturbed configurations, *Shuffled-Text* and *Replaced-Text* in *Text-Summary* prompts. Experiments are conducted with three to five provided examples, and the scores are averaged for each configuration. Across all datasets, each perturbed configuration consistently outperforms the *No-Demo* baseline, and their performances remain relatively close to the *Original* setup. The only noticeable deviation is observed in the SAMSum dataset with the *Replaced-Text* configuration, which, while showing a drop compared to the *Original*, still significantly surpasses the *No-Demo* scores.

mance plateau. This observation highlights the inherent limitation of in-context learning, where adding more examples does not necessarily yield further improvements in performance.

4.2 Probing the Impact of Text-Summary Content Alignment

We introduced controlled perturbations to the examples in *Text-Summary* prompt to probe the efficacy of the content alignment between text and summary. We adopted three distinct experimental setups:

- *Original*: Pairs are taken from the training set, maintaining correct associations. (e.g. $\text{Text}_1 \rightarrow \text{Summ}_1$, $\text{Text}_2 \rightarrow \text{Summ}_2$, $\text{Text}_3 \rightarrow \text{Summ}_3$)
- *Shuffled-Text*: The same texts and summaries are retained, but the order of the texts is shuffled, leading to incorrect associations. (e.g. $\text{Text}_2 \rightarrow \text{Summ}_1$, $\text{Text}_3 \rightarrow \text{Summ}_2$, $\text{Text}_1 \rightarrow \text{Summ}_3$)
- *Replaced-Text*: Only the same summaries are retained. The texts are randomly taken from the training set, making non-related associations. (e.g. $\text{Text}_4 \rightarrow \text{Summ}_1$, $\text{Text}_5 \rightarrow \text{Summ}_2$, $\text{Text}_6 \rightarrow \text{Summ}_3$)

Table 3 unveils some unexpected patterns in the behavior of the perturbations. In the *Shuffled-Text* setup, while the order of texts is rearranged, the associations between texts and summaries are still present, even if they are jumbled. This suggests that if ChatGPT can recognize and realign these associations autonomously, then the perturbation's impact might be minimal. This hypothesis is supported by the observed performance, which remains closely aligned with that of the *Original* setup.

On the other hand, the *Replaced-Text* perturbation presents a more challenging test. It should lead to a significant improvement drop because it involves texts and summaries that are entirely mismatched. However, despite this stark mismatch, performance levels remain comparable to the *Original* setup. The simultaneous presentation of texts and summaries is more critical in ChatGPT's summarization adaptability than the exact content alignment between individual texts and their corresponding summaries. This finding is consistent with the insights provided by Min et al. (2022b), emphasizing the importance of demonstrations' general structure and format over exact input-label alignments in in-context learning.

5 Conclusion

In our exploration of in-context learning, we discerned its profound impact on the text summarization capabilities of ChatGPT. Our findings highlighted that structured demonstrations, pairing texts with summaries, significantly enhance ChatGPT's adaptability, with the structure being more important than exact text-summary content alignment. Nevertheless, our research also unveils the inherent limitations of in-context learning, as evidenced by diminishing improvement with increasing demonstration examples and reduced efficacy in summarizing intricate news texts compared to dialogue summarization. Overall, our research underscores both the potential and the challenges of in-context learning in natural language processing, laying a foundation for future endeavors in text summarization using large language models.

Limitations

While our study offers significant insights into the efficacy of in-context learning for text summarization using ChatGPT, several limitations warrant consideration. Primarily, the exclusive focus on ChatGPT means that the findings may only be par-

tially transferable to other LLMs. Even within the LLM category, each model has unique architecture, training data, and nuances. Thus, generalizing our findings across the board would be premature. Additionally, while we incorporated four datasets from the dialogue and news domains, the absence of datasets from other critical domains means that our findings may need to be more comprehensive for general summarization tasks. Real-world applications often require summarizations of a wide array of text types, including scientific articles, legal documents, and social media posts, to name a few. Our study may need to accurately capture the adaptability of in-context learning for these diverse domains. Future research should expand the scope by integrating more diverse LLMs and datasets from various domains to ensure broader applicability and generalizability of the findings.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022a. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Liang Wang, Nan Yang, and Furu Wei. 2023. Learning to retrieve in-context examples for large language models.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023a. Extractive summarization via chatgpt for faithful summary generation.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023b. Benchmarking large language models for news summarization.

A Appendix

#Exp.	Prompt Type	SAMSum			DialogSum			CNN/DailyMail			XSum		
		R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
0	No-Demo	33.31	12.59	25.60	23.95	8.39	18.62	24.02	9.24	16.65	18.72	5.63	13.04
1	T-S	38.63	15.49	30.18	31.63	11.07	24.75	26.45	9.91	18.36	21.48	6.63	15.34
	T-Only	33.64	12.75	25.99	25.63	9.09	20.02	24.56	9.43	17.06	18.75	5.46	13.05
	S-Only	37.09	13.94	28.82	30.82	10.75	24.19	25.57	9.63	17.74	19.65	5.65	13.85
2	T-S	40.34	16.95	31.72	31.40	11.08	24.60	27.19	10.20	18.86	21.36	6.71	15.25
	T-Only	33.44	12.93	25.81	25.94	9.17	20.24	24.50	9.42	17.09	19.03	5.70	13.29
	S-Only	38.43	14.92	30.16	30.90	10.83	24.28	25.97	9.80	17.97	19.66	5.48	13.77
3	T-S	40.50	17.05	31.86	31.83	11.57	24.98	26.85	10.22	18.72	21.52	6.85	15.38
	T-Only	33.36	12.86	25.72	26.37	9.45	20.66	25.06	9.61	17.36	19.45	6.00	13.69
	S-Only	38.76	15.25	30.42	30.95	11.00	24.36	25.77	9.76	17.88	19.89	5.60	13.96
4	T-S	40.70	17.17	32.07	32.52	11.76	25.67	26.64	10.10	18.62	20.87	6.62	14.88
	T-Only	33.91	13.33	26.22	26.53	9.39	20.71	25.15	9.73	17.56	19.49	5.95	13.65
	S-Only	38.82	15.41	30.39	30.83	10.97	24.25	25.69	9.75	17.77	19.75	5.49	13.83
5	T-S	41.05	17.42	32.40	32.96	12.11	25.93	26.74	10.28	18.65	20.53	6.52	14.59
	T-Only	33.68	13.13	25.82	26.57	9.50	20.77	24.88	9.44	17.28	19.34	6.01	13.58
	S-Only	38.96	15.61	30.64	31.22	11.13	24.55	26.13	9.86	17.91	19.97	5.76	14.10

Table 4: The performance of various summarization prompts on four datasets, categorized by the number of examples (#Exp.) and the type of prompt, including *Text-Summary (T-S)*, *Text-Only* (T-Only), and *Summary-Only* (S-Only). The evaluation is based on Rouge scores (Rouge-1 (R1), Rouge-2 (R2), Rouge-L(RL)) provided by HuggingFace.