Abstract

In this paper we explore efficient summarization techniques for LLMs using 1) parameter-efficient fine-tuning, 2) in-context learning, and 3) Retrieval-Augmented Generation (RAG). We demonstrate that partial fine-tuning of specific model layers, particularly the top layers in encoder blocks, was most effective. Additionally, we found that simpler prompt designs generally outperformed more complex ones. While our investigation of RAG did not yield significant improvements over simple prompts for general summarization, it showed potential for domain-specific applications. Combining these techniques, we achieved ~1% improvement in semantic similarity, approaching state-of-the-art (SOTA) models in semantic evaluation scores.

1. Introduction

In today's data-driven world, it's easy to be overwhelmed by the sheer volume of information. Large Language Models (LLMs) offer a potential solution by condensing information into concise summaries, allowing for the quick grasp of essential insights. This technology has numerous applications, from news consumption to business decision-making, customer feedback analysis, and competitive intelligence.

Summarization is a complex task for machines as they must maintain factual accuracy, coherence, and fluency while being precise. Developing generalized LLMs that can perform summarization well can be costly in time, computational requirements, and the expertise to build the models. Further complicating matters, many cloud hosted LLMs such as ChatGPT, Gemini, and Cohere can raise privacy concerns [1].

Our research aims to address these challenges by exploring and combining three avenues for extreme summarization: efficient parameter fine-tuning, in-context learning, and Retrieval-Augmented Generation (RAG). Previous work has explored these avenues in isolation, while this paper looks at all three avenues holistically. We hypothesize that by combining these techniques, we can develop smaller, self-hosted models that achieve summarization performance comparable to or exceeding that of larger, cloud-based models, while addressing cost and privacy concerns.

2. Related Work

Techniques used for news summarization have experienced evolutionary changes thanks to the deep learning models. These transformer based models, such as BERT, T5, BART and Mistra, are pre-trained to offer strong capabilities to perform various downstream tasks. Tremendous interest has been developed to further enhance these large LLMs to perform a specific task. Fine tuning LLMs is one conventional way to achieve that, which involves adapting the pre-trained models to the small task-specific dataset. However, fully retraining LLMs with all the parameters can be expensive. Hui et al [2] shows that half fine tuning the parameters, as opposed to full fine tuning, offers good balance between existing knowledge learnt and new learning with great computational efficiency. Another popular efficient parameter fine tuning technique is LoRA, developed by Hu el al[3] which freezes the pretrained model parameters and introduces lower rank matrices in each of the layers in the transformer architecture - this results in significantly less trainable parameters, offering an efficient way to adapt the model to new information.

Beyond EFT, Alham, Nurudeen, et al[4] explores various approaches in natural language processing, comparing prompting, in-context learning, and instruction-tuning with respect to the number of labeled samples required. The study found that smaller specialized models can often outperform larger general models with fewer labeled samples, when fine tuned. This paper will expand the work of [4] by investigating prompting, ICL, and instruction tuning so that it may be combined with FT and RAG to determine if smaller models can meet or exceed the performance of larger models.

Additionally, we investigated the impact of incorporating a Retrieval-Augmented Generation (RAG) model on our summarization task which was inspired by the methodologies outlined in Liu et al[5]. The goal was to determine whether integrating a retrieval component would extract and enhance contextual information to improve the quality of summaries produced by Large Language Models (LLMs).

3. Experimental Design

3.1 Dataset

The XSum (Extreme Summarization) dataset is used in this paper as the baseline dataset. Each element of the dataset consists of a BBC article, a human generated summarization, and the BBC article ID. The dataset is split into 204,045 training elements, 11,332 validation and 11,334 test sets [6]. For our experiments, we shuffled the training, validation and test datasets. In the training phase(if applied), we took a varying number of training samples to serve the particular purpose of that experiment. Performance metrics calculations were based on the same 200 examples sampled from the shuffled test dataset (seed = 42) throughout all the experiments for fair comparison.

3.2 Method

We explored three avenues of enhancement strategies: parameter efficient fine tuning, efficient in-context learning, and RAG based summarization. A custom python library standardized data management functions, summarization functions, and evaluation functions were used to ensure consistency throughout all the experiments described in sections XXX. After exploring each avenue, we combined the “best” model found in each avenue and evaluated that model. Pre-trained models used in this paper were the T5 family of encoder-decoder models[7], the Mistral decoder-only model[8], and the BART encoder-decoder model[9].

3.3 Benchmarks and evaluation metrics

This paper extracts the first and last line of the articles as a baseline for summarization, as these sentences often encapsulate the main topic and conclusion of the articles[10]. Baseline performances were presented in Appendix I - table 1. Additionally, we produced some SOTA results using Facebook's BART-large-XSum checkpoint[11,12]as our upper bound to compare our efforts to(Appendix - table 2 and Appendix IV - article 1 & 2 example responses).

For evaluation metrics across all models, content overlap metrics and semantic similarity metrics were used. Content overlap metrics used were ROUGE[13] and BLUE[14] scores and semantic similarity metrics used were BERTScore[15] and Vector Similarity scores[16]. It should be noted that these metrics are the most commonly used metrics to quantify summarization performance, but they do not necessarily capture faithfulness and fluidity[17].

Therefore this paper also includes qualitative discussion on some output summaries to complete the evaluation.

4. Experiments, Results and Discussions

4.1 Parameter Efficient Fine Tuning(EFT)

We studied two efficient parameter fine tuning techniques 1) partial Full Fine Tuning(FT) and 2) LoRA. With that objective, we conducted a sequence of experiments to understand the models incrementally. Some initial experiments for learning purpose and to provide context for the EFT strategies are shared in Appendix III - Table 1&2.

Key results

Model	**# of trainable params**	**Rouge1**	**Rouge 2**	**RougeLsum**	**BLUE**	**BERT-F1**	**Vector Similarity**
T5-base - FT top layer(encoder blocks) + cross-attention layers(decoder block)	111,988,992	0.27	0.08	0.21	0.04	0.88	0.54
T5-base - FT top layer(encoder blocks)	83,668,224	**0.28**	**0.08**	**0.22**	**0.05**	0.88	**0.56**
LoRA - rank 4, alpha =4	442,368	0.18	0.02	0.13	0.01	0.85	0.48
LoRA - rank 16, alpha =16	1,769,472	0.19	0.03	0.14	0.01	0.86	0.48

Analysis

1) Partial Fine Tuning

With similar intuition as Kaplun et al [18], we hypothesize that not all layers are equally effective for fine tuning and their effectiveness depends on the specific downstream task.intuitively, lower layers captured more linguistic features, and the top layers were often fine-tuned to adapt to the model's final predictions. Some research [19] also suggested that last(top) layer fine-tuning alone can effectively promote fairness in deep neural networks.

To adapt for our specific task of highly abstractive short summarization, we focused on fine tuning the top layers in the encoder blocks, hypothesizing that those are crucial to adjust the encoder outputs on new data and the task for extreme summarization. We also combined that with FT the cross-attention layers in the decoder blocks to see if refining cross attentions would help.

Compared to the benchmarks, the partial FT T5-base pretrained model showed promising results. FT T5-small seemed not complex enough to adjust for new data and tasks, and could even give some wrongful information such as the example in Appendix IV (row 4 in Table 1 & 2). T5 large was still computational prohibitive with our resource constraints so we skipped.

Selectively fine tuning certain layers, especially the top layer in the encoder blocks showed the most effectiveness. With only ⅓ of trainable parameters of the full model, it was able to score almost on par with the full FT T5 model and not far from the BART-large-XSum except for the BLEU score (see Key Results Table above vs. benchmarks in Appendix I). The summary outputs were precise ,abstract and fluid, matching well in style with the references when we manually inspected some summary outputs shown in Appendix IV (row 5 in Table 1&2).

2) LoRA

Introduced by Hu et al [20], LoRA has gained traction in efficiently retraining LLMs and adapting to new datasets and tasks. In our experiments, we focused on varying the number of ranks. Generally, lower rank means less details learnt and updates the parameters. Whitehouse et al [21] found that LoRA is competitive with full fine-tuning when trained with high quantities of data. Following the work by them and also the optimal rank reported by Che et al[22], we use rank 4 and 16 in our experiments along with alpha equals to the rank to keep the scalar impact constant in both cases.

Compared to benchmarks, the quantitative metrics were not meaningfully better. We found that even with a substantial training sample, low rank seemed to constrain the learning on some key features. For example, rank 4 produced very similar summaries to those from the T5 pretrained models(i.e.containing a few extractive sentences but losing features could cause inaccurate info.). When rank was increased to 16, the output summaries started to get shorter and more precise(i.e.moving to the direction of extreme summarization). Examples are again shown in Appendix IV (row 6,7 in Table 1&2).

Overall, partial FT with the ‘right’ layers was more effective than LoRA in our experiments, which showed significant improvement to the benchmarks while balancing output quality and computational burdens.

4.2 - Efficient in-context learning

In-context learning is the ability of a machine learning model to learn and adapt to new tasks or information based on the context provided alongside the task. The summarization capabilities of the T5-small, T5-base, T5-large models, and Mistral models were evaluated using three distinct techniques:simple prompts, template-based prompts, and multi-shot learning prompts (see Appendix V). Each experiment involved generating summaries across the XSum dataset as described in section 31. Each experiment was evaluated using the metrics described in section 3.3, with detailed results presented in Appendix VI. Representative outputs from the experiments can be found in Appendix VII.

Analysis

Baseline: Both simple and template-based prompts outperformed the lower bound baseline model (as detailed in Appendix I - Table 1), while multi-shot learning prompts did not. However, none of the in-context learning modifications reached the performance of the upper bound baseline model (Appendix I - Table 2).

Cross-experiment comparisons: For simple, template, and multi-shot learning prompt design, variations across the various metrics within a particular model were not substantial as seen by the average and standard deviation of the various metrics in Appendix X. This indicates that alterations of the prompt did not substantially affect the quality of the model's output.

Prompt effectiveness: Simpler prompts outperformed both template-based summarization and multi-shot learning across the tested models. Multi-shot learning showed substantial lower scores across the board than simple prompt design and template based summarization for both the T5 model family and the Mistral model. This behavior can be attributed to context windows, lack of pre-training and fine-tuning of the models on the multi-shot templates and limited capacity of the model architecture.[23]

Model comparison: The Mistral model performs better across semantic similarity scores than the T5 models, but has lower content overlap scores than the T5. This indicates that the Mistral model shows strong performance in capturing the underlying meaning or semantics of the text (as reflected by higher vector similarity and BERTScore), the Mistral model does not perform as well in terms of directly matching specific content or surface-level text overlap when compared to T5 models. This divergence in performance can be attributed to the fundamental differences in the architectures and training objectives of these two model types. T5 models are inherently designed to prioritize factual accuracy, while Mistral models are optimized for nuanced understanding and fluent generation [7].

Qualitative: The T5 models (base, small, large) generally focus on the key details of the incident: the involvement of a minibus, the tragic death of Bethany Jones, and the sentencing of the driver. However, the Mistral model outputs tend to include additional context or narrative scope, such as the aftermath for Sarah Johnson and her work with a charity. The differences also reflect how each model balances factual reporting with additional contextual information, influencing the summary's clarity and informativeness.

4.3 - Retrieval Augmented Generation-Based Summarization

This experiment uses a two-part system, a retrieval component and a LLM that takes the retrieval context to generate a summary.

For the retrieval system, we leveraged a text splitter to divide articles into chunks with a defined ‘chunk_size’ and ‘chunk_overlap’ to ensure each chunk has a coherent segment of text while maintaining context continuity. The document chunks are fed into a FAISS vector store along with a specific article id, and we used the sentence transformer embeddings to represent the text chunks and to retrieve based on semantic similarity. As for the query used for comparing semantic similarity, we looked into using different sentences in each article because they typically follow an inverted pyramid structure. To ensure the documents retrieved are from the same article, we assign each article a specific id and insert the id as metadata for the chunks in the vector store.

As for the LLM step, the documents retrieved were fed into two different LLMs, the t5-base and mistral model, for generating summaries. The LLMs were not fine-tuned during this experiment to reduce computational complexity, and the results were. Detailed results can be seen in Appendix VIII.

Analysis

From the experiments, the best results came from splitting the articles into a maximum chunk size of 200, no overlap, and 10 documents retrieved. From the evaluation metrics, we noticed that using the first sentence or both sentences (first and last) as queries consistently led to the retrieval of more relevant document chunks. These strategies were more effective since the first sentence usually provides a broad overview and main point of the article. When combined with the last sentence, the retrieval system also includes the concluding thought and potentially a summary of the article. On the other hand, using only the last sentence often resulted in retrieving only very specific segments of the document, and it may not always reflect the key points needed for a well-rounded summary.

Baseline: Overall, the Mistral model results showed significant improvement over the baseline sentences. Similarly to the query retrieval, using the both first and last sentences would gather the main context, but using the retrieval system and LLMs result in a better summarized output of the article compared to extracting sentences as our result because the retrieval system gains more context to summarize. However, compared to the state of the art BART model, the RAG system is not up to par, particularly because the BART model was trained on the XSum dataset.

Cross-experiment comparison: In comparing the performance of the Mistral model with T5, the mistral model showed more promising results for summarization. The mistral model appears to be more oriented towards the task of summarization since it demonstrates a better ability to synthesize information from the retrieved chunks and generate a cohesive and coherent summary. The summaries generated by Mistral showed better cohesion and fluency while the T5 generated summaries looked to be more extractive of individual sentences and sometimes were not fluent.

The difference in approach between the two models can be attributed to abstractive vs extractive summarization. Mistral's ability to generate novel, task-oriented text shows a stronger abstractive capability, which is essential for summarization with the XSum dataset where the goal is to create a concise representation of the document without necessarily using the exact wording of the source material. In contrast, T5's tendency towards extractive summarization, where it heavily relies on the phrasing from the input text, can limit its ability to generate summaries that are both concise and representative of the articles' relevant content.

4.4 - “Best combination” - Efficient fine tuning of T5 model and prompt3

We decided that the “best” combination to evaluate was the EFT T5-base model combined with prompt3. We found that EFT produced the best results when applied to the encoder, and as the Mistral model is decoder only, that precluded its use. Further, we found that the RAG summarization did perform better than the using the model without RAG.

Detailed results can be seen in Appendix X, with the “best” model performing marginally better than the EFT T5-base model on semantic similarity scores, but worse on content overlap scores. The “best” model performed close to the SOTA model on semantic scores, but substantially worse on content overlap scores.

5.0 Conclusions and further work

In our exploration of efficient summarization techniques, we found that partial fine-tuning, particularly targeting specific layers in the encoder and decoder blocks, led to significant improvements in model performance. Our experiments highlight that simple prompts generally outperform template-based and multi-shot prompts, suggesting a need for simpler yet effective prompt engineering for smaller models, or that smaller models need to be partially trained on more complex prompts. We also found that the Mistral model generally outperformed the T5 model on semantic similarities scores and produced more fluid output, but is an order of magnitude larger than the largest T5 model tested. Combining prompts and EFT showed that simpler models can approach SOTA models semantic similarity performance, but not in content overlap performance.

Future work should look at differences between semantic similarity and content overlap performance of the combined model. Further, while the RAG based summarization did not provide noticeable improvements in summarization scores, additional research in the area could be undertaken.

References

All code can be found in: https://github.com/sleighton2022/datasci266-final-project

[1]] Kang, S. Y., Lee, H. J., & Kim, J. H. (2024). Text in-context learning: Challenges, recent advances, and prospects for machine learning applications. *Fundamental Research*, *4*(3), 493-508.
[2] https://arxiv.org/html/2404.18466v1
[3] https://arxiv.org/abs/2106.09685
[4] Alham, Nurudeen, et al. "Fine-Tuning, Prompting, In-Context Learning and Instruction-Tuning: How Many Labelled Samples Do We Need?" *arXiv preprint arXiv:2402.12819*. 2024.
[5] arXiv:2403.19889
[6] Narayan, S., Cohen, S. B., & Lapata, M. (2018). XSum: Extreme Summarization. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 546–557. https://huggingface.co/datasets/EdinburghNLP/xsum
[7] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, *21*(140), 1-67.
[8] https://arxiv.org/abs/2310.06825
[9] https://arxiv.org/abs/1910.13461
[10] https://arxiv.org/abs/2204.01849
[11] https://arxiv.org/abs/1910.13461
[12] https://huggingface.co/facebook/bart-large-xsum
[13] https://aclanthology.org/W04-1013.pdf
[14] https://aclanthology.org/P02-1040.pdf
[15] https://arxiv.org/abs/1904.09675
[16] https://arxiv.org/pdf/2204.01632
[17] https://aclanthology.org/2022.konvens-1.8.pdf
[18] https://openreview.net/forum?id=sOHVDPqoUJ
[19] https://openreview.net/pdf?id=wrmynnlrTI
[20] https://arxiv.org/abs/2106.09685
[21] https://arxiv.org/pdf/2311.08572v2
[22] https://arxiv.org/abs/2405.02710
[23] https://arxiv.org/pdf/2301.13688

RAG
References:
[1] https://arxiv.org/abs/2402.13446
[2] https://arxiv.org/abs/2403.19889v1
[3] https://arxiv.org/pdf/2310.06825

Appendix I: Baseline

Table 1 : extractive baseline

Model	rouge1	rouge2	rougeL	BLEU	BERT Score	Vector Similarity
baseline-first sentence	0.17	0.02	0.13	0.01	0.47	0.43
baseline-last sentence	0.13	0.01	0.10	0.01	0.44	0.28
baseline-both sentence	0.17	0.02	0.12	0.01	0.49	0.46

Table 2: BART-large-XSum (STOA)

Model	rouge1	rouge2	rougeL	BLEU	BERT Score	Vector Similarity
BART-large-XSum	0.37	0.16	0.29	0.11	0.90	0.65

Appendix II: Example XSum articles and summarizations

Article 1:

Input Article	Sarah Johnson was one of 21 women heading to Liverpool when their minibus was hit by a lorry on the M62. Her friend Bethany Jones, 18, was killed while Ms Johnson and several others were badly hurt. Minibus driver James Johnson was jailed for more than six years for causing Bethany's death, in April 2013. Ms Johnson, who broke her shoulder, back and pelvis, said the help she received from a charity while in hospital led her to want to support others. Speaking publicly for the first time about the crash, Ms Johnson described how everyone was "excited and giddy" for the hen party. "To me the impact was just a massive explosion," she said. "I thought the bus had blown up. "I remember the bus dropping on its side. The next thing, I woke up on the roadside so I'd actually come out of the window." Ms Johnson was taken to Leeds General Infirmary where she, along with Bethany's sister Amy Firth, underwent major surgery and spent time in intensive care. Whilst she was there she got support from charity Day One, which helps victims of major trauma. She said: "It's absolutely fantastic. "It supports people by giving benefit advice, legal advice and peer support such as me and Amy, who have been in similar situations and who are now helping other people who've suffered from major trauma." Ms Johnson said the crash had made her realise how lucky she had been. "Beth can't complain, she's not here," she added. "We just have to be grateful for what we've got."
Reference	**A woman who was seriously hurt in a fatal hen party motorway crash is now helping other major trauma victims rebuild their lives.**

Article 2:

Input Article	Two-year-old Lane Thomas Graves had been playing in the sand near the resort's Seven Seas Lagoon when he was dragged underwater by the creature. His parents and older sister had been visiting the Grand Floridian resort in June 2016 from the state of Nebraska. The lighthouse has been installed near to where the attack occurred. Wildlife officials classified the killing as a predatory attack, saying the boy did nothing to provoke the alligator. "He was in the water not more than ankle deep," the Florida Fish and Wildlife Conservation Commission said in a report, describing how the boy had been gathering water for a sandcastle. His father, Matt Graves, jumped in the water to try to pry open the creature's mouth, but "the alligator thrashed and broke Matt's grasp and went under the water," according to the report. A Disney spokesperson said they hoped the monument would spread awareness for the Lane Thomas Foundation, which also uses the lighthouse as its logo. Who is liable for alligator boy's death? "The lighthouse sculpture has been installed to help spread awareness of the Lane Thomas Foundation, which was established to provide assistance and support to families whose children need organ transplants," Walt Disney World said in a statement. After the death, Disney was criticised for not having posted signs warning of the danger along the man-made lagoon, which borders Magic Kingdom. Public notices have now been added to the area, Florida media report. The Lane family announced a month after the boy's death that they would not sue Disney, and would instead "solely be focused on the future health of our family".
Reference	**Walt Disney World has unveiled a lighthouse memorial for a young boy who was killed by an alligator while on holiday at the Florida theme park.**

Appendix III: Experiments and learning preceding efficient fine tuning

Using the pre-trained and available checkpoint models , we developed some fundamental understanding. We learnt that A- the pretrained T5 models, despite different sizes, did not produce meaningfully different summaries without fine tuning (they shared the pattern of having all multiple extractive sentences, shuffled in orders and parts). B- the direct benefit from BART-large-CNN on the XSum task was limited, as the label summaries were in distinct styles(multiple sentences longer summary vs. extreme summarization). C- BART-large-XSum provided STOA results and the full FT T5 also provided strong results in terms of both quantitative and qualitative aspects. More detailed results and key learning takeaways are shown in Table 1 & 2 below:

Table 1 - results and key learning takeaways on pretrained models

Model	**Rouge1**	**Rouge 2**	**RougeLsum**	**BLUE**	**BERT-F1**	**Vector Similarity**
T5-small	0.17	0.02	0.12	0.01	0.85	0.45
T-5 base	0.18	0.02	0.13	0.01	0.85	0.49
T-5 large	0.19	0.03	0.13	0.01	0.85	0.51
BART -large-CNN	0.19	0.02	0.13	0.01	0.86	0.49

- Model sizes did not affect the generated summaries meaningfully. Although quantitative scores improved as the model got bigger, when we qualitatively inspected some random summary outputs, we found that all the pre-trained T5 models producing the summaries with similar patterns across the three model sizes -they tended to have the pattern of taking the first few sentences with some shuffling or parts. This is reasonable with news articles which typically have key messages upfront, however, the summary did not look very fluid by squeezing a few sentences together, compared to the label reference that was more precise. This indicated that adapting the pretrained models for our specific downstream task was necessary.

- BART-large-CNN had very similar evaluation scores across board to those of the T5-base model. This made intuitive sense, as summary references in the CNN news usually contained multiple long sentences, similar to what a pre-train T5 would generate. This suggested the knowledge transfer from that dataset was not particularly suitable for extreme summary. Fine tuning was needed to tailor for short summaries.
- BART-large-XSum, not surprisingly, gave very impressive results as it's fully FT on the XSum dataset atop a powerful BART model. Not only the quantitative scores were meaningfully higher, but also the summaris read by us manually were more abstract and very close to the label references. This served as the upper bound for us to check our experiments against

Table 2 - results and key learning takeaways on Full FT models

Model	**# of training sample**	**Rouge1**	**Rouge 2**	**RougeLsum**	**BLUE**	**BERT-F1**	**Vector Similarity**
T5-base Full FT	300	0.25	0.06	0.18	0.03	0.87	0.54
T5-base Full FT	5000	0.27	0.07	0.20	0.04	0.88	0.54
T5-base Full FT	20000	**0.28**	**0.09**	**0.22**	**0.05**	**0.89**	**0.57**

- When fully training the T-5 small model coupled with small training sample size (200 and 5000), despite improved evaluation scores some summary outputs appeared overfitting , either containing wrongful information or nonsense(e.g. “dragged underwater by a sea urchin” or “by a sandcastle” as opposed to “an alligator” in the article).
- When fully training the T-5 base model coupled with small training size, results were not significantly better than those from the pre-trained models when we inspected them - they appeared to be mostly direct parts from the original article similar to what the pre-trained T5 would produce,i.e. quite extractive and somewhat lack of fluency
- Only when we increased training sample size substantially to 20000 on the T-5 base model, we noticed meaningful improvement in the generated summaries, which were more abstract, concise, faithful and fluid, although it took a much longer time to fully fine tune the full model.

Appendix IV: Efficient fine tuning example responses

Table 1: Responses for Appendix II: Article 1

Model	Model Renerated Summary
BART-large-XSum(((STOA)	**A woman who was on a hen party when her minibus was hit by a lorry has spoken of the "massive explosion" of the crash that killed her friend.**
Pre-trained T5-large	minibus driver jailed for six years for causing death of friend in minibus crash . Sarah Johnson, 21, was one of 21 women in minibus hit by lorry on m62 . friend Bethany Jones, 18, was killed while several others were badly hurt . minibus driver james johnson jailed for more than six years for causing crash .
Full FT T5-base	An 18-year-old woman who was killed when her minibus was hit by a lorry has spoken out about her hen party in Liverpool.

T5-small FT last layer in each encoder block + Decoder -second layers(cross-attention) in each decoder block	A woman has been jailed for causing her death ***[poor faithfulness and fluidity]*** by a lorry on a hen party in Liverpool after a lorry hit a minibus.
t5-base - FT top layer(encoder blocks)	An 18-year-old woman who was killed when her minibus was hit by a lorry has spoken out about her hen party in Liverpool.
LoRA(rank = 16)	Sarah Johnson was one of 21 women heading to Liverpool when their minibus was hit by a lorry . her friend Bethany Jones, 18, was killed while Ms Johnson and several others were badly hurt .
LoRA(rank = 4)	Sarah Johnson was one of 21 women heading to Liverpool when their minibus was hit by a lorry . her friend Bethany Jones, 18, was killed while Ms Johnson and several others were badly hurt . minibus driver James Johnson was jailed for more than six years for causing her friend's death .***[inaccurate info.]***

Table 2: Responses for Appendix II: Article 2

Model	Model Renerated Summary
BART-large-XSum (STOA)	**A lighthouse has been built in memory of a toddler who was killed by an alligator while on holiday in the Florida resort town of Bradenton.**
Pre-trained T5-large	two-year-old Lane Thomas Graves was dragged underwater by an alligator . wildlife officials classified the killing as a predatory attack . the boy did nothing to provoke the alligator, officials say .
Full FT T5-base	A Florida boy has been killed by an alligator in a sandcastle at the Grand Floridian Resort, officials say.
T5-small FT last layer in each encoder block + Decoder -second layers(cross-attention) in each decoder block	Two-year-old Lane Thomas Graves was dragged underwater by a sandcastle ***[nonsense]*** in the Florida state of Florida.
t5-base - FT top layer(encoder blocks)	Florida's Grand Floridian resort has been declared a "predatory attack" after a boy was dragged underwater by an alligator.
LoRA(rank = 16)	Lane Thomas Graves, 2, was playing in the sand near the resort's Seven Seas Lagoon when he was dragged underwater
LoRA(rank = 4)	Lane Thomas Graves, 2, was playing in the sand near the resort's seven seas lagoon when he was dragged underwater by the creature . his parents and older sister had been visiting the resort in June 2016 from the state of Nebraska .

Appendix V: In-context learning prompts

Simple prompts:

Prompt	Prompt content
prompt1	"Summarize this article: {document}"
prompt2	"What are the key points of this article: {document}"
prompt3	"Summarize this article for a 5th grader: {document}"
prompt4	"Write a summary of this article in 50 words: {document}"
prompt5	"Summarize the article in 3 bullet points: {document}"

Template based summarization prompts:

Prompt	Prompt content
template1	Input: Article: {document} Task: Summarize the above article. Output: Summary:
template2	Input: Article: {document} Task: Extract and summarize the key points from the article. Output: Key Points Summary:
template3	Input: Article: {document} Task: Summarize this article for a 5th grader. Output: Summary (50 words):
template4	Input: Article: {document} Task: Summarize the article in approximately 50 words. Output: Summary (50 words):
template5	Input: Article: {document} Task: Summarize the article in bullet points. Output: Summary: - bullet_point_1 - bullet_point_2 - bullet_point_3

Multi-shot learning prompts

Prompt	Prompt content
learning1	"Document: " + learning_dataset["document"][0] + "Summary:" + \\ learning_dataset["summary"][0] + " Summarize the following {document}"
learning2	"Document: " + learning_dataset["document"][0] + "Summary:" + \\ learning_dataset["summary"][0] + \\ "Document: " + learning_dataset["document"][1] + "Summary:" + \\ learning_dataset["summary"][1] + " Summarize the following {document}"

learning_dataset = dataset_manager.load_sampled_dataset(dataset_label="train")

Appendix VI: In-context learning quantitative results

Simple prompts: t5-base

Experiment	rouge1	rouge2	rougeL	bleu	bertscore	vector_similarity
t5-base-prompt1	0.19	0.02	0.13	0.02	0.50	0.51
t5-base-prompt2	0.20	0.03	0.13	0.02	0.51	0.49
t5-base-prompt3	0.19	0.03	0.13	0.01	0.51	0.53

t5-base-prompt4	0.20	0.02	0.13	0.01	0.50	0.51
t5-base-prompt5	0.19	0.02	0.13	0.02	0.50	0.50
Mean	0.19	0.03	0.13	0.02	0.50	0.51
Std deviation	0.01	0.00	0.00	0.00	0.00	0.02

Simple prompts: t5-small

Experiment	rouge1	rouge2	rougeL	bleu	bertscore	vector_similarity
t5-small-prompt1	0.19	0.02	0.13	0.01	0.50	0.43
t5-small-prompt2	0.19	0.02	0.13	0.01	0.49	0.44
t5-small-prompt3	0.19	0.02	0.13	0.01	0.49	0.42
t5-small-prompt4	0.18	0.02	0.12	0.01	0.50	0.45
t5-small-prompt5	0.18	0.03	0.13	0.01	0.49	0.43
Mean	0.19	0.02	0.13	0.01	0.49	0.43
Std deviation	0.01	0.00	0.00	0.00	0.00	0.01

Simple prompts: t5-large

Experiment	rouge1	rouge2	rougeL	bleu	bertscore	vector_similarity
t5-large-prompt1	0.20	0.03	0.14	0.02	0.51	0.49
t5-large-prompt2	0.20	0.03	0.14	0.02	0.51	0.47
t5-large-prompt3	0.20	0.03	0.13	0.01	0.51	0.50
t5-large-prompt4	0.19	0.03	0.13	0.02	0.50	0.48
t5-large-prompt5	0.20	0.03	0.13	0.02	0.50	0.49
Mean	0.20	0.03	0.14	0.02	0.51	0.49
Std deviation	0.01	0.00	0.00	0.00	0.00	0.01

Simple prompts: mistral

Experiment	rouge1	rouge2	rougeL	bleu	bertscore	vector_similarity
mistral-prompt1	0.08	0.02	0.06	0.01	0.47	0.60
mistral-prompt2	0.08	0.02	0.05	0.00	0.47	0.60
mistral-prompt3	0.08	0.02	0.06	0.01	0.48	0.58
mistral-prompt4	0.10	0.02	0.07	0.01	0.48	0.56
mistral-prompt5	0.09	0.02	0.06	0.01	0.47	0.59
Mean	0.09	0.02	0.06	0.01	0.47	0.58
Std deviation	0.01	0.00	0.01	0.00	0.01	0.02

Template based summarization: t5-base

Experiment	rouge1	rouge2	rougeL	bleu	bertscore	vector_similarity
t5-base-template1	0.19	0.02	0.13	0.01	0.50	0.49
t5-base-template2	0.20	0.03	0.13	0.01	0.50	0.48
t5-base-template3	0.19	0.03	0.12	0.01	0.50	0.48
t5-base-template4	0.19	0.02	0.12	0.01	0.50	0.49

t5-base-template5	0.18	0.03	0.11	0.01	0.49	0.47
Mean	0.19	0.02	0.12	0.01	0.50	0.48
Std deviation	0.01	0.00	0.01	0.00	0.00	0.01

Template based summarization: t5-small

Experiment	rouge1	rouge2	rougeL	bleu	bertscore	vector_similarity
t5-small-template1	0.18	0.02	0.13	0.01	0.49	0.42
t5-small-template2	0.17	0.02	0.12	0.01	0.50	0.44
t5-small-template3	0.17	0.02	0.12	0.01	0.49	0.41
t5-small-template4	0.17	0.02	0.12	0.01	0.49	0.42
t5-small-template5	0.18	0.02	0.12	0.01	0.49	0.42
Mean	0.17	0.02	0.12	0.01	0.49	0.42
Std deviation	0.00	0.00	0.00	0.00	0.00	0.01

Template based summarization: t5-large

Experiment	rouge1	rouge2	rougeL	bleu	bertscore	vector_similarity
t5-large-template1	0.20	0.03	0.14	0.02	0.50	0.49
t5-large-template2	0.21	0.03	0.14	0.02	0.51	0.49
t5-large-template3	0.21	0.03	0.15	0.02	0.50	0.48
t5-large-template4	0.21	0.03	0.15	0.02	0.50	0.49
t5-large-template5	0.21	0.03	0.14	0.01	0.51	0.49
Mean	0.21	0.03	0.14	0.02	0.51	0.49
Std deviation	0.00	0.00	0.00	0.00	0.00	0.00

Template based summarization: Mistral

Experiment	rouge1	rouge2	rougeL	bleu	bertscore	vector_similarity
mistral-template1	0.08	0.02	0.06	0.00	0.46	0.53
mistral-template2	0.07	0.02	0.05	0.00	0.45	0.50
mistral-template3	0.08	0.02	0.06	0.01	0.47	0.49
mistral-template4	0.08	0.02	0.06	0.01	0.46	0.52
mistral-template5	0.05	0.01	0.04	0.00	0.40	0.48
Mean	0.07	0.02	0.05	0.00	0.45	0.50
Std deviation	0.01	0.00	0.01	0.00	0.03	0.02

Multi-shot learning: T5-large

Experiment	rouge1	rouge2	rougeL	bleu	bertscore	vector_similarity
t5-large-learning1	0.07	0.00	0.04	0.01	0.39	0.05
t5-large-learning2	0.05	0.00	0.03	0.01	0.40	0.07
Mean	0.06	0.00	0.04	0.01	0.40	0.06
Std deviation	0.01	0.00	0.01	0.00	0.01	0.02

Multi-shot learning: Mistral

Experiment	rouge1	rouge2	rougeL	bleu	bertscore	vector_similarity
mistral-learning1	0.02	0.01	0.02	0.00	0.38	0.07
mistral-learning2	0.03	0.01	0.02	0.00	0.38	0.07
Mean	0.02	0.01	0.02	0.00	0.38	0.07
Std deviation	0.00	0.00	0.00	0.00	0.00	0.00

Appendix VII: Representative outputs for in-context learning for qualitative analysis

Responses for Appendix II: Article 1

t5-base-prompt1	21 women were heading to Liverpool when their minibus was hit by a lorry . their friend, 18-year-old Bethany Jones, was killed in the crash . minibus driver James Johnson was jailed for more than six years for causing her death
t5-small-prompt1	her friend Bethany Jones, 18, was killed while her minibus was hit by a lorry . the driver was jailed for more than six years for causing the crash . she said the help she received from a charity led her to want to support others
t5-large-prompt1	Sarah Johnson was one of 21 women on a minibus that was hit by a lorry on the M62. Her friend Bethany Jones, 18, was killed and several others were badly hurt.
mistral-prompt1	Sarah Johnson was one of 21 women on a minibus that was hit by a lorry on the M62 in 2013. Her friend Bethany Jones was killed in the crash, and Sarah is now working as a support worker for Day One,
t5-base-template1	her friend Bethany Jones, 18, was killed while her minibus was hit by a lorry . the driver was jailed for more than six years for causing the crash . she said the help she received from a charity led her to want to support others
t5-small-template1	minibus driver jailed for more than six years for causing death of friend . Sarah Johnson broke her shoulder, back and pelvis in the crash . Ms Johnson says crash made her realise how lucky she was
t5-large-template1	minibus driver jailed for more than six years for causing death of friend . Sarah Johnson broke her shoulder, back and pelvis in the crash . Ms Johnson says crash made her realise how lucky she was
mistral-template1	Sarah Johnson was one of 21 women on a minibus that was hit by a lorry on the M62 in 2013. Her friend Bethany Jones was killed in the crash, and Sarah is now working as a support worker for Day On
t5-large-learning1	['academies are directly funded by central government and head teachers have more freedom over admissions and the way the school works . it is a significant development in the continued divergence of schools systems on either side of the border. the Welsh Government will get extra cash to match the money for english schools to extend the school day, but it can spend it on any

	devolved policy area .']
mistral-learning1	'Ms Johnson, who is now working as a support worker for Day One, said she wanted to give back to the charity that had helped her.\nShe said: "It\'s something I\'m passionate about.\n"It\'s'

Appendix VIII: RAG results

Top results:

model	query	chunk size	overlap size	k size	rouge1	rouge2	rougeL	BLEU	BERT Score	Vector Score
mistral	both	200	0	10.00	0.22	0.05	0.16	0.03	0.54	0.54
mistral	first	200	0	15.00	0.22	0.04	0.15	0.02	0.53	0.54
mistral	both	200	50	10.00	0.21	0.04	0.15	0.02	0.53	0.53
mistral	both	200	0	15.00	0.21	0.04	0.15	0.02	0.53	0.52
mistral	first	200	50	10.00	0.21	0.04	0.15	0.02	0.53	0.52

All results:

model	query	chunk size	overlap size	k size	rouge1	rouge2	rougeL	BLEU	BERTScore F1	Vector Similarity	Avg Score
mistral	first & last	200	0	10	0.22	0.05	0.16	0.03	0.54	0.54	0.31
mistral	first sentence	200	0	15	0.22	0.04	0.15	0.02	0.53	0.54	0.30
mistral	first & last	200	50	10	0.21	0.04	0.15	0.02	0.53	0.53	0.30
mistral	first & last	200	0	15	0.21	0.04	0.15	0.02	0.53	0.52	0.30
mistral	first sentence	200	50	10	0.21	0.04	0.15	0.02	0.53	0.52	0.30
mistral	first sentence	200	50	15	0.21	0.04	0.15	0.02	0.53	0.52	0.29
mistral	first sentence	200	0	10	0.21	0.04	0.14	0.02	0.53	0.52	0.29
mistral	first & last	200	50	15	0.21	0.04	0.15	0.02	0.52	0.52	0.29
mistral	first sentence	200	0	5	0.21	0.04	0.15	0.02	0.52	0.52	0.29
mistral	first sentence	100	50	15	0.21	0.04	0.15	0.02	0.52	0.51	0.29
mistral	first sentence	200	50	5	0.21	0.04	0.14	0.02	0.52	0.51	0.29
mistral	first & last	100	0	15	0.21	0.04	0.15	0.02	0.53	0.50	0.29
mistral	first & last	100	0	10	0.21	0.04	0.15	0.02	0.53	0.50	0.29
mistral	first sentence	100	0	15	0.21	0.04	0.14	0.02	0.52	0.50	0.29
mistral	first sentence	100	0	10	0.21	0.04	0.15	0.02	0.52	0.49	0.29
mistral	first & last	200	50	5	0.20	0.03	0.14	0.02	0.52	0.51	0.28
mistral	first sentence	200	0	3	0.21	0.03	0.15	0.02	0.52	0.50	0.28
mistral	first	100	50	10	0.20	0.03	0.14	0.02	0.52	0.50	0.28

	sentence										
mistral	first & last	100	50	15	0.20	0.03	0.14	0.02	0.52	0.49	0.28
mistral	first & last	100	50	10	0.20	0.03	0.14	0.02	0.52	0.49	0.28
mistral	first & last	200	50	3	0.20	0.04	0.14	0.02	0.52	0.48	0.28
mistral	first sentence	200	50	3	0.20	0.03	0.14	0.02	0.51	0.49	0.28
t5-base	first & last	200	0	15	0.20	0.03	0.14	0.02	0.51	0.48	0.27
t5-base	first & last	200	0	15	0.20	0.03	0.14	0.02	0.51	0.47	0.27
t5-base	first & last	200	50	10	0.20	0.03	0.13	0.02	0.51	0.47	0.27
t5-base	first & last	200	50	15	0.20	0.03	0.13	0.02	0.51	0.48	0.27
t5-base	first sentence	200	50	15	0.19	0.03	0.13	0.02	0.51	0.47	0.27
t5-base	first & last	200	0	10	0.20	0.03	0.14	0.02	0.51	0.46	0.27
t5-base	first sentence	200	0	15	0.19	0.03	0.13	0.02	0.50	0.47	0.27
t5-base	first sentence	200	50	10	0.19	0.03	0.13	0.02	0.50	0.47	0.27

Appendix IX - Representative output of RAG-based summarization

Example Reference:
A woman who was seriously hurt in a fatal hen party motorway crash is now helping other major trauma victims rebuild their lives.

Retrieval Strategy - First Sentence:
Document: Sarah Johnson was one of 21 women heading to Liverpool when their minibus was hit by a lorry on the M62. Her friend Bethany Jones, 18, was killed while Ms Johnson and several others were badly hurt.
Document: Minibus driver James Johnson was jailed for more than six years for causing Bethany's death, in April 2013.
Document: Ms Johnson was taken to Leeds General Infirmary where she, along with Bethany's sister Amy Firth, underwent major surgery and spent time in intensive care.
Document: Ms Johnson said the crash had made her realise how lucky she had been.
"Beth can't complain, she's not here," she added. "We just have to be grateful for what we've got."
Document: Speaking publicly for the first time about the crash, Ms Johnson described how everyone was "excited and giddy" for the hen party.

Retrieval Strategy - Last Sentence:
Document: Ms Johnson said the crash had made her realise how lucky she had been.
"Beth can't complain, she's not here," she added. "We just have to be grateful for what we've got."
Document: helping other people who've suffered from major trauma."
Document: "I remember the bus dropping on its side. The next thing, I woke up on the roadside so I'd actually come out of the window."
Document: Ms Johnson, who broke her shoulder, back and pelvis, said the help she received from a charity while in hospital led her to want to support others.

Document: "To me the impact was just a massive explosion," she said. "I thought the bus had blown up.

Retrieval Strategy - Both Sentences:
Document: Sarah Johnson was one of 21 women heading to Liverpool when their minibus was hit by a lorry on the M62.
Her friend Bethany Jones, 18, was killed while Ms Johnson and several others were badly hurt.
Document: Ms Johnson said the crash had made her realise how lucky she had been.
"Beth can't complain, she's not here," she added. "We just have to be grateful for what we've got."
Document: Speaking publicly for the first time about the crash, Ms Johnson described how everyone was "excited and giddy" for the hen party.
Document: Ms Johnson was taken to Leeds General Infirmary where she, along with Bethany's sister Amy Firth, underwent major surgery and spent time in intensive care.
Document: Ms Johnson, who broke her shoulder, back and pelvis, said the help she received from a charity while in hospital led her to want to support others.

RAG Mistral Results:
In an interview, Sarah Johnson shared her experience of surviving a fatal car accident that killed her friend Bethany Jones during a hen party trip. She expressed gratitude for being alive but also mourned the loss of her friend. While recovering from injuries at Leeds General Infirmary, Ms. Johnson discovered a charity called Day One, which provided support to victims of major trauma like herself.

RAG T5 Results:
21 women were heading to Liverpool when their minibus was hit by a lorry . their friend, 18-year-old Bethany Jones, was killed in the crash . the minibus driver was jailed for more than six years for causing her friend's death .

EXAMPLE: 5
Reference: Blackpool have signed Bournemouth goalkeeper Ryan Allsop and midfielder Callum Cooke from Middlesbrough on season-long loans.
Mistral: The two young English footballers, Allsop and Cooke, both had successful careers before joining Blackpool for the upcoming season.
T5: Allsop made his premier league debut in November 2015 . the 25-year-old has spent time on loan at coventry and Wycombe .

Appendix X - “Best combination” results

Quantitative results

Experiment	rouge1	rouge2	rougeL	bleu	bertscore	vector_similarity
t5-eft-prompt	0.21	0.04	0.15	0.01	0.90	0.59

Representative outputs

sample1-labeled	A woman who was seriously hurt in a fatal hen party motorway crash is now helping other major trauma victims rebuild their lives.
Sample1-generated	Sarah Johnson was one of 21 women injured when their minibus was hit by a lorry on the M62 in Liverpool . her friend, 18-year-old Bethany Jones, was killed while she and

	several others were badly hurt . a minibus driver was jailed for more than six years for causing her friend's death .
sample2-labeled	A Tudor manor house has reopened following a £2.2m makeover.
sample2-generated	A total of 1,400 tickets have been sold out for the opening weekend of the new Bramall Hall in Stockport, Greater Manchester . the manor dates back to the reign of William the Conqueror when he bestowed the lands on one of his followers, the first Baron of Dunham Massey .
sample3-labeled	Walt Disney World has unveiled a lighthouse memorial for a young boy who was killed by an alligator while on holiday at the Florida theme park.
sample3-generated	A two-year-old boy has been dragged underwater by an alligator . his parents and sister had been visiting a resort in the state of Nebraska . officials say the boy did nothing to provoke the alligator .