Cross-Domain Evaluation of Large Language Models for Abstractive Text Summarization: An Empirical Perspective

Walid Mohamed Aly[1], Taysir Hassan A. Soliman[2], Amr Mohamed AbdelAziz[3]
Information Systems Department-Faculty of Computers and Information, Assiut University, Assiut, Egypt, 71515[1,2]
Information Systems Department-Faculty of Computers and Artificial Intelligence,
Beni-Suef University, Beni-Suef, Egypt, 62111[3]

***Abstract*—Large Language Models (LLMs) have demonstrated remarkable capabilities in generating human-like text; however, their effectiveness in abstractive summarization across diverse domains remains underexplored. This study conducts a comprehensive evaluation of six open source LLMs across four datasets: CNN / Daily Mail and NewsRoom (news), SAMSum (dialogue) and ArXiv (scientific) using zero shot and in-context learning techniques. Performance was assessed using ROUGE and BERTScore metrics, and inference time was measured to examine the trade-off between accuracy and efficiency. For long documents, a sentence-based chunking strategy is introduced to overcome context limitations. Results reveal that in-context learning consistently enhances summarization quality, and chunking improves performance on long scientific texts. The model performance varies according to architecture, scale, prompt design, and dataset characteristics. The qualitative analysis further demonstrates that the top-performing models produce summaries that are coherent, informative, and contextually aligned with human-written references, despite occasional lexical divergence or factual omissions. These findings provide practical insights into designing instruction-based summarization systems using open-source LLMs.**

***Keywords*—Large language models; natural language processing; automatic text summarization; prompt engineering; summarization evaluation**

I. Introduction

With the continued explosion of digital content from online news, scientific research, and conversational platforms, the demand for automated methods of condensed textual information has increased. Automatic Text Summarization (**ATS**) has become an essential tool in Natural Language Processing (NLP), reducing cognitive load and enabling users to quickly grasp key information quickly [1]. In recent years, the task of summarization has gained significance with the rise of Large Language Models (LLMs), which demonstrate remarkable fluency in generating human-like text across a variety of domains [2], [3].

Efforts in text summarization have evolved from traditional heuristic-based methods [4] to more sophisticated approaches to cover different types of text summarization including generic [5], domain-aware [6], multi-document [7], multimodal [8], extractive [9] and abstractive text summarization [10], [11]. Various techniques are used in text summarization to enhance the produced summary, including machine learning approaches [12], deep learning-based approaches [13], [14]. The advent of transformer-based models [15], [16] further advanced the quality of summarization by improving semantic understanding and text generation.

A major shift in NLP occurred with the development of instruction-tuned LLMs, which are pretrained on vast corpora and require no task-specific fine-tuning. Instead, they rely on prompt-based interactions, enabling paradigms such as zero-shot and in-context learning [17], [2]. Despite these advancements, there is limited understanding of how instruction-tuned, open-source LLMs perform across various domains of summarization—especially in contexts such as scientific documents and informal dialogues.

Prompt Engineering is a sophisticated AI engineering methodology [18]. This involves augmenting LLMs by giving them customized cues and modifying the source text to produce the intended result. Prompt engineering is crucial in LLMs because it is essential for unlocking the full potential of such models. In addition, prompt engineering uses prior knowledge and the logical reasoning of the input to influence the outputs generated by the Models [19]. Wide-ranging techniques have been developed as a result of recent notable advancements in the field of prompt engineering [20]. The range of these developments includes basic techniques and more advanced strategies intended to manage challenging jobs [21].

This work contributes to the NLP field by offering domain-aware benchmarking of LLMs under prompt-based conditions. This study also explores the relationship between model architecture, parameter count, and performance across different textual formats. Ultimately, the study provides practical insights for researchers and developers seeking to leverage LLMs for summarization tasks in real-world scenarios without relying on extensive fine-tuning pipelines. Our primary contributions are summarized as follows:

- An evaluation of six open-source LLMs across four benchmark datasets: CNN/DailyMail, NewsRoom, SAMSum, and ArXiv, providing insights into their cross-domain summarization abilities.
- A comparative assessment of prompting strategies, including Zero-Shot Learning (ZSL) and In-Context Learning (ICL), to evaluate the effect of prompt design, context length, and number of demonstrations on model performance.

- An empirical analysis of model scale and architecture to examine the impact of parameter size and design on summarization quality across different domains.
- A dedicated investigation of long-document summarization, introducing a chunking strategy to mitigate the limitations imposed by context window size, and evaluating its influence on summary coherence and quality.
- A detailed efficiency and cost analysis measuring inference time for each model, highlighting trade-offs between performance and deployment feasibility.

The remainder of this paper is organized as follows: Section II introduces related works on text summarization and LLMs. Section III describes in detail the experimental setup and methods, Section IV interprets the results, Section V provides Discussion, Finally, Section VI provides a concise conclusion and future work.

II. Literature Review

Text summarization research has evolved considerably , with two primary paradigms: extractive and abstractive summarization [4]. Extractive methods select sentences or phrases directly from the source document to form summaries [9], while abstractive approaches generate novel sentences by interpreting the core meaning of the input [10].

Early summarization systems, such as the Lead-3 baseline, extracted the first few sentences to generate reliable summaries for news datasets [22]. More advanced neural models treated sentence selection as a classification problem using hierarchical RNNs [23] or applied encoder-decoder architectures with attention mechanisms to improve sentence representation and relevance scoring [24]. The introduction of transformer architectures [25] further enhanced abstractive summarization by capturing long-range dependencies in text. Fine-tuning transformer-based models like BERT [26] for summarization tasks demonstrated strong performance in both extractive [27] and abstractive [28] settings.

State-of-the-art models like BART [29] and PEGASUS [30] have since advanced abstractive summarization. BART's bidirectional encoding and autoregressive decoding enables robust generation, while PEGASUS's gap-sentence generation objective improves content selection and coherence in generated summaries. Variants like Hi-BART further integrate hierarchical encoders to improve the structural understanding of long documents [31].

With the rise of LLMs, prompt-based summarization has gained traction. Instruction-tuned LLMs, such as GPT-3, LLaMA, and Mistral, demonstrate strong zero-shot and few-shot capabilities without the need for task-specific training [2]. Studies like [32] and [33] evaluated such models on standard datasets (e.g., CNN/DailyMail, XSum), demonstrating superior performance for models like text-davinci-003. Other research expanded this assessment to clinical summarization tasks and reported mixed results across radiology reports, patient dialogues, and medical questions [34].

Prompt Engineering is a sophisticated AI engineering methodology [18]. This involves augmenting LLMs by giving them customized cues and modifying the source text to produce the intended result. Prompt engineering is crucial in LLMs because it is essential for unlocking the full potential of such models. In addition, prompt engineering uses prior knowledge and the logical reasoning of the input to influence the outputs generated by the Models [19]. Wide-ranging techniques have been developed as a result of recent notable advancements in the field of prompt engineering [20].

Despite these advancements, long-document summarization remains a key challenge. Standard LLMs struggle to process entire scientific articles due to limited input windows. Techniques such as hierarchical summarization [7] and chunking [35] have been proposed to segment inputs and generate localized summaries prior to merging them. Although effective, these methods often introduce trade-offs between coherence and inference efficiency.

Our study advances previous works by systematically evaluating six open-source LLMs across diverse domains using both zero-shot and in-context prompting strategies. In contrast to studies focused on single-domain tasks or isolated techniques, we offer a unified assessment of summary quality and computational efficiency. In addition, we introduce and empirically evaluate a sentence-level chunking strategy for long document summarization that has not been evaluated systematically in text summarization across instruction-tuned models.

III. Methodology and Experimental Setup

This study evaluates the performance of open-source LLMs in text summarization across multiple domains. To ensure reproducibility, this section outlines the datasets, models, and inference settings used.

A. Datasets

We employed four benchmark datasets encompassing news, dialogue, and scientific texts, which were selected to enable comprehensive cross-domain evaluation. Dataset statistics are summarized in Table I.

1) CNN/Dailymail: The CNN/DailyMail dataset [36] comprises over 300 K news articles from CNN and the Daily Mail, which are commonly used for training and evaluating summarization models.

2) Cornell NEWSROOM: The NEWSROOM dataset [37] includes summaries from 38 news publishers collected via search and social media metadata (1998–2017), encompassing diverse summarization strategies across extractive and abstractive styles.

3) SAMSum corpus: SAMSum [38] contains 16 K annotated chat dialogues with human-written abstractive summaries, which were designed to support dialogue summarization in informal communication settings.

4) ArXiv dataset: The ArXiv dataset [39] features full-text research papers across disciplines such as physics, biology, and computer science, accompanied by human-generated abstracts. The long document structure poses significant challenges for LLM summarization.

In our experiments, we sampled 2 K test instances per dataset, except SAMSum, where all test samples were used.

For few-shot prompting, examples were drawn from each dataset's training set.

TABLE I. STATISTICS OF THE USED DATASETS (K=THOUSANDS)

Dataset	Domain	Documents			Sum Len
		Train	Valid	Test	(words)
CNN/DM	News	287K	13K	11K	52
NEWSROOM	News	995K	108K	108K	26
SAMSum	Dialogue	14K	818	819	–
ArXiv	Scientific	203K	6K	6K	220

B. Prompt Engineering Techniques

As introduced in Section I, prompt engineering involves crafting input instructions to guide LLM behavior. This study applies two primary strategies—**zero-shot learning (ZSL)** and **in-context learning (ICL)**—to investigate how prompt formats affect summarization performance across different models and datasets.

1) Zero-Shot Learning (ZSL) [40]: ZSL is the simplest form of prompting, where a model receives only natural language instructions without any prior examples. The model relies entirely on internal knowledge to perform the task. Formally, a ZSL prompt can be expressed as follows:

$$P = f_{\text{prompt}}(\text{TD}, x_{\text{test}}) \tag{1}$$

where:

- TD is the task description,
- x_{test} is the test input,
- f_{prompt} is the function that transforms these into a natural language prompt.

2) In-Context Learning (ICL) [41]: In ICL, the model observes a few input-output pairs (demonstrations) directly in the prompt and then predicts the output for a new test input. This process can be described as follows:

$$P(y_{\text{test}} \mid x_{\text{test}}, D_k) = \text{LLM}([x_1, y_1, \ldots, x_k, y_k, x_{\text{test}}]) \tag{2}$$

where $D_k = \{(x_1, y_1), (x_2, y_2), \ldots, (x_k, y_k)\}$ denotes k demonstrations. The LLM uses these in-context examples to condition its prediction for x_{test} without updating model parameters.

C. Large Language Models Selection

This study focuses on autoregressive language models—specifically decoder-only architectures—which have become the dominant paradigm in open-source LLM development (Table II). These models generate text by sequentially predicting the next token based on preceding tokens; thus, they are well-suited for text generation tasks, such as summarization. Unlike sequence-to-sequence models, which use an encoder-decoder framework and were originally designed for translation [42], autoregressive models rely solely on a decoder and are optimized for unsupervised learning and generative tasks [43].

We evaluated six decoder-only models, including three from the LLaMA-2 family [44]—LLaMA-2-7B-chat, LLaMA-2-13B-chat, and LLaMA-2-70B-chat—to investigate the effect of scaling on summarization performance. These models are widely adopted due to their open-source availability and strong instruction-tuning capabilities.

In addition, we include Mistral-7B-Instruct-v0.1 [45] and Gemma-7B-it [46], both lightweight instruction-tuned models with 7B parameters. To explore architectural variations, we incorporated Mixtral-8x7B-Instruct-v0.1 [47], a 47B-parameter Mixture-of-Experts (MoE) model that activates only relevant subnetworks during inference, thereby reducing computational costs [48].

This curated selection enabled both intra-scale comparisons between the 7B models and cross-scale evaluations involving larger models. By analyzing these models under uniform evaluation settings, we assess how instruction tuning, architectural innovations, and scaling affect summarization quality and efficiency.

Proprietary models were excluded to ensure reproducibility and mitigate biases associated with closed-source data or opaque updates. The experiments were conducted on two NVIDIA A100 GPUs using mixed-precision (fp16), model parallelism, and quantization techniques to optimize memory usage.

TABLE II. CHARACTERISTICS OF DIFFERENT LLMS UTILIZED IN OUR WORK

Model Name	# Parameters	Context-Length
Llama-2-7b-chat	7B	4K
gemma-7b-it	7B	8K
Mistral-7B-Instruct-v0.1	7B	8K
Llama-2-13b-chat	13B	4K
Mixtral-8x7B-Instruct-v0.1	47B	32K
Llama-2-70b-chat	70B	4K

D. Long Document Processing via Chunking Strategy

Summarizing long documents, such as scientific papers from the ArXiv dataset, presents challenges due to the limited context window of LLMs. Initially, we addressed this by trimming articles to fit the model's maximum context length. Although effective for compatibility, this approach risks omitting important details.

Although extended-context models offer potential solutions, they are not universally available or efficient. To address this issue, we adopt a chunking strategy that divides documents into semantically coherent segments—such as paragraphs or fixed-length chunks. Each chunk is summarized independently, and the resulting summaries are then merged and refined to form a coherent global summary [49].

To improve chunk relevance, we used the Natural Language Toolkit (NLTK)[1] to segment documents into sentences. This

[1]https://www.nltk.org/

ensures semantic integrity in each chunk and ensures that key content is retained during summarization. This process allows for effective handling of long documents while aligning with the input constraints of the models.

E. Evaluation Metrics

The evaluation of summarization quality requires robust metrics that capture both surface-level similarity and deeper semantic alignment. Accordingly, we employ two metric categories: **Word Overlap** and **Semantic Similarity**.

1) Word overlap metrics: These metrics compare the token overlap between the generated and reference summaries. We adopt ROUGE [50], a standard suite for this purpose.

ROUGE-N measures the overlap of n-grams (contiguous sequences of n words) between the system and reference summaries, which indicates lexical similarity.

ROUGE-L evaluates the Longest Common Subsequence (LCS) between summaries, reflecting structural and sequential alignments.

Higher ROUGE scores suggest greater textual overlap and structural resemblance to reference summaries.

2) Semantic similarity metrics: These metrics assess how well a summary preserves the meaning of the original text. We use BERTScore [51], which leverages contextual embeddings from BERT [26] to compute token-level semantic similarity.

Unlike ROUGE, BERTScore captures nuanced paraphrasing and long-range dependencies by evaluating contextual token embeddings. In addition, it excels when n-gram methods are short, particularly in abstractive summarization scenarios.

F. Inference Parameters

This section outlines the inference settings and prompt configurations used in our experiments. Given that the underlying mechanisms of ZSL and ICL were previously detailed, we focused on their practical deployment and tuning.

For ZSL, we employed diverse task-specific prompts tailored to each dataset. In ICL, we varied the number of demonstrations: 1, 3, 5, and 7 to explore the trade-off between contextual guidance and token budget, guided by previous studies [41], [35].

Temperature is a critical generation hyperparameter that controls randomness during decoding [52]. Lower values yield more deterministic and factual outputs, whereas higher values encourage diverse but potentially less accurate generations. Before the full evaluation, temperatures of 0.1, 0.5, and 0.9 were tested. The results in Table III indicate that 0.1 consistently produced the highest ROUGE scores, especially ROUGE-L; thus, it was adopted for all final evaluations.

To leverage the potential of LLMs with ZSL, the different models are guided using a variety of prompts, aiming to assess the results across different prompts to fully utilize the models' capabilities. The prompts are customized based on the diversity of each dataset to provide summaries and assess the output of each model against the reference summaries. The prompts for zero-shot prompting are available in Table IV.

TABLE III. PERFORMANCE ACROSS ALL DATASETS AND TEMPERATURE VALUES. P=PRECISION, R=RECALL, F1=F1-SCORE, RL=ROUGE-L. BOLD VALUES INDICATE BEST PERFORMANCE PER METRIC GROUP

Temp	Metric	Dataset	P/R1	R/R2	F1/RL
0.1	BERT	CNNDM	88.76	86.13	87.41
		NewsRoom	86.99	**87.67**	**87.31**
		SAMSum	89.31	**91.48**	**90.37**
		ArXiv	**86.86**	80.09	83.33
	ROUGE	CNNDM	37.44	**16.42**	**24.53**
		NewsRoom	**25.38**	**8.82**	**19.96**
		SAMSum	**39.35**	**14.61**	**30.70**
		ArXiv	**49.74**	**32.47**	**33.98**
0.5	BERT	CNNDM	86.35	**88.03**	87.16
		NewsRoom	**87.11**	87.30	87.28
		SAMSum	**90.54**	88.76	89.62
		ArXiv	85.67	81.85	83.70
	ROUGE	CNNDM	**38.38**	14.73	24.49
		NewsRoom	24.42	8.19	18.41
		SAMSum	39.08	14.13	30.25
		ArXiv	40.20	12.92	22.09
0.9	BERT	CNNDM	88.40	86.13	87.24
		NewsRoom	87.02	87.40	87.19
		SAMSum	90.30	88.51	89.38
		ArXiv	85.67	**82.03**	**83.80**
	ROUGE	CNNDM	36.87	13.03	23.64
		NewsRoom	24.28	80.01	18.91
		SAMSum	39.34	14.36	30.53
		ArXiv	39.29	12.19	21.61

IV. RESULTS

This section presents the performance of the evaluated LLMs in producing accurate and coherent summaries. All generated outputs were cleaned to eliminate noise and formatting artifacts before evaluation.

A. Comparison of Prompting Techniques

We evaluated the impact of prompting strategies—ZSL and ICL—on summarization performance across different models and datasets. This analysis highlights how prompting format influences the quality and consistency of the generated summaries.

1) ZSL Results: Multiple zero-shot prompts were tested to instruct the models to generate concise summaries without training examples. Fig. 1 shows the comparative performance of LLMs using BERTScore F1 and ROUGE-1 metrics on the four datasets. Subfigure (a) presents BERTScore F1; (b) presents ROUGE-1.

a) CNN/DM Results in ZSL: [2] Table III summarizes the prompt types tested. As shown in Table V, the results can be explored in two ways:

(i) Length-constrained prompts (e.g., Prompts 3, 4, 5), which restrict summary length based on the dataset's average (see Table I); and

(ii) Structured prompts, which provide directive phrasing to guide model output without specifying length.

[2]Only top-performing prompts are shown. Full list available at: https://anonymous.4open.science/r/TextSummarizationCode-5BCB/README.md

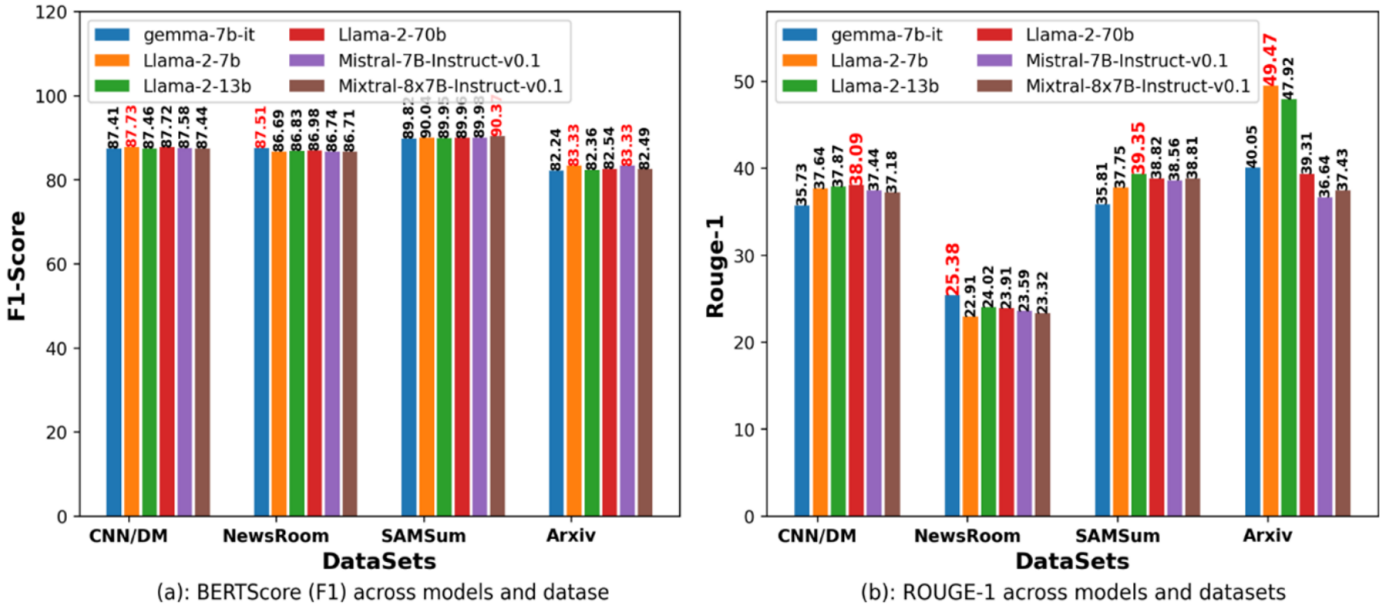


Fig. 1. Performance of LLMs across datasets in the zero-shot setting. X-axis: datasets; Y-axis: ROUGE-1 and BERTScore F1 scores.

These variations help assess how instruction format impacts zero-shot summarization performance.

Comparing models on the same prompt shows that Mistral-7B-Instruct-v0.1 achieved the best ROUGE score on Prompt#1. For BERTScore, Llama-2-70B-chat leads in Recall and F1, while Llama-2-13B-chat scores highest in Precision. Prompt#2 results favor Llama-2-13B-chat across all metrics except Recall, where Llama-2-7B-chat excels. Prompts 3 and 4, being more descriptive, yield varied performance: Mistral-7B and Llama-2-70B lead in ROUGE, while BERTScore is shared among gemma-7b-it, Llama-2 variants, and Mistral-7B, highlighting that some models capture semantic meaning well even if lexical similarity is lower.

Comparing prompts within each model shows that gemma-7b-it performed best on Prompt#3, with top scores in ROUGE-1, ROUGE-L, and BERTScore F1. The Llama-2-7B and Llama-2-13B chats achieved the highest metrics with Prompts#5 and #2 respectively. For Mistral-7B-Instruct-v0.1 and Mixtral-8x7B-Instruct-v0.1, Prompts#3 and #2 are most effective, with Mixtral also showing strong F1 with Prompt#5. These findings underscore the critical role of prompt design in shaping LLM performance.

b) NewsRoom results in ZSL: The NewsRoom dataset includes both extractive and abstractive summaries. In this study, we focus solely on articles with abstractive summaries to align with the goal of evaluating the LLMs' ability to generate coherent, human-like outputs [37]. This provides a more rigorous benchmark for testing generative capabilities. Table VI presents the ZSL results for all models and prompts.

In BERTScore, gemma-7b-it consistently performed well, achieving the highest Precision (86.99), Recall (87.67), and F1 (87.31) with Prompt#1. Llama-2-70B-chat closely follows, particularly in terms of F1 score (86.97), while Mistral-7B-Instruct-v0.1 excels in Recall (88.17) and Precision across several prompts.

For ROUGE, gemma-7b-it again leads with top scores in ROUGE-1 (25.38), ROUGE-2 (8.82), and ROUGE-L (19.96) using Prompt#1. Llama-2-13B-chat and Llama-2-70B-chat demonstrated competitive ROUGE scores across prompts, while Mistral-7B-Instruct-v0.1 achieved its highest ROUGE-1 (23.59) also with Prompt#1.

Overall, Prompts#1 and #2 yield the highest ROUGE and BERTScore results for most models, likely due to their clear and direct instructions. Slight variations in other prompts result in marginally lower performance, illustrating how prompt phrasing influences model output.

In summary, gemma-7b-it performed best on BERTScore, Llama-2-70B-chat demonstrated robust performance across prompts, and Mistral-7B-Instruct-v0.1 proved versatile with strong scores in both lexical and semantic metrics, particularly under well-structured prompt settings.

c) SAMSum results in ZSL: The proposed SAMSum dataset presents challenges for LLMs because of its conversational structure. Table VII compares model outputs for different prompts. The models achieved higher ROUGE scores on news datasets, reflecting the distinct nature of dialogue summarization.

The BERTScore and ROUGE metrics generally agreed. Mixtral-8x7B-Instruct-v0.1 achieved the highest BERT F1 with Prompts#2 and #4, while Llama-2-13B-chat yielded the top ROUGE-1 scores with the same prompts, reinforcing metric consistency.

Simpler prompts (e.g., Prompts#1 and #2) lead to higher ROUGE scores across models, suggesting that clear, concise instructions are more effective. The descriptive prompts (#3–#5) show mixed performance, indicating that overly detailed instructions did not always enhance the results.

Overall, Mistral-7B-Instruct-v0.1 demonstrated robust performance across prompts, demonstrating adaptability for sum-

TABLE IV. COMPLETE LIST OF ALL PROMPTS USED ACROSS DATASETS (P=PROMPT). FULL TEXT PRESERVED EXACTLY

ID	Prompt Text
CNN/DailyMail Prompts	
P#1	Write a concise and comprehensive summary of this news article.
P#2	Provide an abstract of this news article in a direct and concise summary.
P#3	I want you to act as a text summarizer to help me create a concise summary of the provided text. The summary can be up to 3 sentences in length, expressing the key points and concepts written in the original text without adding your interpretations.
P#4	I want you to act as a text summarizer to help me create a concise summary of the provided text. The summary can be up to 2 sentences in length, expressing the key points and concepts written in the original text without adding your interpretations.
P#5	Please summarize the following news article in an informative extractive summary with two sentences.
NewsRoom Prompts	
P#1	Summarize this news article in one sentence.
P#2	Write a concise and comprehensive summary of this news article in one sentence.
P#3	I want you to act as a text summarizer to help me create a concise summary of the following article.
P#4	Provide an abstract of this news article in a direct and concise summary in a few words.
P#5	You are a helpful assistant. Please summarize the following text in one sentence.
SAMSum Prompts	
P#1	Summarize the following dialogue.
P#2	Summarize the following dialogue in a few words.
P#3	Summarize the following dialogue into an abstractive summary without any explanation.
P#4	In short, what's going on in this conversation?
P#5	Summarize the following conversation.
P#6	Summarize this conversation in one or two sentences.
ArXiv Prompts	
P#1	Provide an abstract of the following research article.
P#2	Summarize the main points and findings of the scientific paper.
P#3	I want you to act as a research paper summarizer to get an abstract for this research paper.
P#4	Given this research article, create a TLDR to be used as a formal abstract for this paper.

marizing dialogues. Llama-2-13B-chat also performed well, particularly with direct prompts, indicating its suitability for concise summarization.

d) ArXiv results in ZSL: The ArXiv dataset includes scientific papers and human-written abstracts; thus, it is ideal for testing abstractive summarization. Because of the context-length limitations, documents were trimmed before inference. Table VIII presents model performance across prompts.

Prompt#2 consistently yields high ROUGE scores, especially for Llama-2-13B-chat and Mistral-7B-Instruct-v0.1, demonstrating its effectiveness in guiding models to capture key information. Prompt#3 performs well in F1/RL scores, indicating the value of explicit summarization instructions. Prompt#4 balances BERT and ROUGE scores, helping models generate concise and relevant abstracts. Overall, the ROUGE scores varied more than the BERT scores, highlighting differences in lexical alignment across models.

Llama-2-7B-chat performed best overall, particularly on Prompt#2, confirming its strength in abstractive summarization. Mistral-7B-Instruct-v0.1 excels with Prompt#3, demonstrating consistency in semantic capture. In contrast, gemma-7b-it outperformed the other models, which suggests its limitations in handling longform texts. The variability across prompts and metrics underscores the complexity of long-document summarization and the impact of input trimming on model comprehension.

2) ICL Results: ICL uses example-summary pairs to guide models in generating new outputs. Based on the prompt formulation in Section III, we evaluate ICL performance on each dataset.

a) CNN/DM results in ICL: Table IX presents results with varying numbers of demonstrations. BERT metrics are relatively stable across shot counts, with minor gains from more examples. ROUGE metrics show greater sensitivity: Mistral-7B, Mixtral-8x7B, and Llama-2-70B benefit notably as shots increase.

Llama-2-70B-chat achieves the highest overall scores with 7-shot prompts, and it excels in both BERT and ROUGE. Mixtral-8x7B also performs well. Gemma-7b-it shows limited gains from more demonstrations, indicating constraints in leveraging ICL effectively.

b) NewsRoom results in ICL: Table X summarizes the ICL results obtained on the NewsRoom dataset. **Mixtral-8x7B-Instruct-v0.1** shows the most consistent improvement, particularly with 5-shot prompts, achieving strong performance in both ROUGE and BERTScore metrics. The ROUGE scores varied more across shot counts than the BERT metrics. The Llama family models performed competitively, especially in 7-shot settings, although they were slightly less consistent across fewer examples. Overall, increasing the number of demonstrations generally enhanced the performance across the models.

c) SAMSum results With ICL: when working with ICL on the SAMSum dataset, Table XI showed that **Mixtral-8x7B-Instruct-v0.1** got the best results in terms of BERTScore and Rouge score with seven examples to learn from, which means that it learned the context of the dialogues better than others. The more examples given to the model, the more the result improved, Table XI shows that the results improved when the number of demonstrations in most models was increased.

d) ArXiv results in ICL: Table XII shows ICL results obtained on the ArXiv dataset. Across varying numbers of demonstrations, the **Llama** models consistently outperformed the other models in terms of both ROUGE and BERT metrics. The performance generally improved as the number of in-context examples increased, reflecting better adaptation to the complex summarization task. However, due to input trimming (required by context length constraints), variations in model performance are more pronounced, as some models may struggle to retain contextual coherence after truncation.

e) Overall comparison (ICL vs. ZSL): ICL consistently enhances LLM performance across datasets compared to ZSL by providing contextual demonstrations that improve task understanding. Our experiments demonstrate that increasing the number of in-context examples leads to steady performance

TABLE V. PERFORMANCE OF LLMS ON CNN/DM DATASET USING ZSL. BOLD INDICATES BEST PERFORMANCE PER METRIC GROUP

Prompts		Prompt #1		Prompt #2		Prompt #3		Prompt #4	
Metric		BERT	ROUGE	BERT	ROUGE	BERT	ROUGE	BERT	ROUGE
gemma-7b-it									
	P/R1	85.73	31.69	85.99	34.52	**88.76**	36.66	88.23	35.26
	R/R2	88.24	12.37	87.51	13.04	86.13	13.62	**85.91**	12.38
	F1/RL	86.96	20.17	86.72	24.99	87.41	24.48	87.04	23.74
Llama-2-7b-chat									
	P/R1	84.84	28.95	86.24	32.29	86.73	36.41	86.98	37.19
	R/R2	88.28	11.66	**88.23**	11.13	88.11	14.59	87.88	14.21
	F1/RL	86.51	19.84	87.22	19.47	87.41	23.55	87.71	23.51
Llama-2-13b-chat									
	P/R1	**85.98**	32.13	**86.89**	**37.87**	87.41	37.21	87.85	37.48
	R/R2	88.17	13.45	88.06	**17.07**	87.56	13.57	86.71	13.36
	F1/RL	87.05	21.58	**87.46**	**25.99**	87.47	24.01	87.39	24.15
Llama-2-70b-chat									
	P/R1	85.95	29.44	86.81	35.31	87.39	37.08	87.92	**37.76**
	R/R2	**88.53**	11.02	88.07	15.12	87.96	14.19	87.56	**14.82**
	F1/RL	**87.21**	18.39	87.42	25.29	**87.66**	24.03	**87.72**	**25.08**
Mistral-7B-Instruct-v0.1									
	P/R1	85.89	**37.02**	86.57	37.44	86.82	**37.44**	86.91	36.78
	R/R2	88.05	**16.50**	87.88	15.14	**88.38**	**16.42**	87.83	14.56
	F1/RL	86.94	**24.35**	87.21	24.08	87.58	**24.53**	87.36	23.85
Mixtral-8x7B-Instruct-v0.1									
	P/R1	85.38	33.88	86.33	37.18	86.96	35.54	86.91	35.48
	R/R2	88.25	13.94	88.06	15.13	87.25	13.04	86.95	12.28
	F1/RL	86.77	22.12	87.17	24.93	87.09	22.79	86.91	22.02

TABLE VI. PERFORMANCE OF LLMS ON NEWSROOM DATASET USING ZSL

Prompts		Prompt #1		Prompt #2		Prompt #3		Prompt #4	
Metrics		BERT	ROUGE	BERT	ROUGE	BERT	ROUGE	BERT	ROUGE
gemma-7b-it									
	P/R1	**86.99**	**25.38**	86.72	24.28	85.63	21.85	85.93	22.89
	R/R2	**87.67**	**8.82**	87.71	8.14	87.75	6.99	87.74	7.59
	F1/RL	**87.31**	**19.96**	87.19	18.81	86.65	16.44	86.81	17.18
Llama-2-7b-chat									
	P/R1	85.19	***22.78***	84.71	21.69	84.51	20.13	85.38	22.15
	R/R2	88.23	7.81	88.29	7.45	88.23	6.72	88.04	6.87
	F1/RL	***86.67***	17.07	86.44	15.97	86.31	14.73	86.67	16.15
Llama-2-13b-chat									
	P/R1	85.42	***24.02***	85.07	22.99	84.88	21.38	85.69	21.76
	R/R2	88.34	8.19	88.41	8.05	88.41	7.02	87.72	6.78
	F1/RL	***86.83***	17.89	86.68	17.17	86.58	15.61	86.68	16.26
Llama-2-70b-chat									
	P/R1	85.71	***23.71***	85.37	23.47	84.75	21.17	85.40	21.81
	R/R2	88.31	8.25	88.39	8.04	88.25	7.16	88.11	7.17
	F1/RL	***86.97***	18.19	86.84	17.51	86.44	15.59	86.71	16.03
Mistral-7B-Instruct-v0.1									
	P/R1	85.40	***23.59***	84.42	20.59	83.43	17.33	83.56	17.44
	R/R2	88.17	9.10	88.09	7.52	88.01	6.01	88.56	6.02
	F1/RL	***86.74***	18.34	86.18	15.61	85.63	12.84	85.69	12.95
Mixtral-8x7B-Instruct-v0.1									
	P/R1	85.39	***23.32***	84.57	21.62	83.96	19.12	83.98	19.52
	R/R2	88.11	7.72	88.17	7.06	88.07	6.37	88.86	7.52
	F1/RL	***86.71***	17.21	86.32	15.83	85.95	13.97	86.01	14.23

TABLE VII. PERFORMANCE OF LLMS ON THE SAMSUM DATASET WITH ZSL

Prompts		Prompt #1		Prompt #2		Prompt #3		Prompt #4	
Metrics		BERT	ROUGE	BERT	ROUGE	BERT	ROUGE	BERT	ROUGE
gemma-7b-it									
	P/R1	**86.24**	**33.16**	88.21	34.31	86.41	32.05	**88.76**	35.23
	R/R2	**91.05**	**11.48**	90.10	11.70	90.68	10.29	90.62	12.14
	F1/RL	**88.58**	**25.25**	89.02	26.02	88.48	24.29	89.66	26.72
Llama-2-7b-chat									
	P/R1	85.12	***32.72***	89.48	37.75	86.94	34.22	88.76	37.73
	R/R2	90.51	10.80	90.64	13.35	90.55	11.31	90.62	13.47
	F1/RL	***87.72***	24.86	90.04	29.90	88.69	26.09	89.66	30.00
Llama-2-13b-chat									
	P/R1	86.09	***34.48***	89.23	***38.96***	88.29	36.66	89.16	**39.35**
	R/R2	90.55	12.41	90.72	14.22	91.27	12.41	90.78	14.61
	F1/RL	***88.26***	26.70	89.95	30.47	89.74	27.68	89.88	30.70
Llama-2-70b-chat									
	P/R1	87.02	***35.11***	89.09	38.31	88.56	***37.40***	89.09	38.82
	R/R2	91.14	12.32	90.14	14.21	91.34	13.47	90.89	**14.97**
	F1/RL	***89.02***	26.98	89.59	30.20	***89.91***	29.03	89.96	**30.59**
Mistral-7B-Instruct-v0.1									
	P/R1	87.01	***34.24***	89.11	38.56	87.64	34.85	89.07	38.55
	R/R2	91.51	11.91	90.92	13.94	91.12	12.03	90.71	14.31
	F1/RL	***89.19***	26.51	89.98	30.26	89.33	26.64	89.86	30.38
Mixtral-8x7B-Instruct-v0.1									
	P/R1	86.83	***33.67***	89.31	38.81	87.65	35.07	**89.38**	37.97
	R/R2	91.18	10.78	**91.48**	13.89	91.19	11.27	91.14	13.58
	F1/RL	***88.93***	25.28	**90.37**	30.16	89.37	26.85	***90.23***	29.60

TABLE VIII. PERFORMANCE OF LLMS ON THE ARXIV DATASET WITH ZSL

Prompts		Prompt #1		Prompt #2		Prompt #3		Prompt #4	
Metrics		BERT	ROUGE	BERT	ROUGE	BERT	ROUGE	BERT	ROUGE
gemma-7b-it									
	P/R1	84.27	25.72	84.51	40.05	86.14	24.73	86.09	22.52
	R/R2	77.81	15.76	79.46	25.64	78.69	13.45	78.45	12.04
	F1/RL	80.89	18.61	81.89	27.81	82.24	18.05	82.09	16.59
Llama-2-7b-chat									
	P/R1	83.86	***39.61***	83.92	**49.74**	**86.86**	***38.25***	84.73	***36.19***
	R/R2	79.56	21.52	**80.41**	**32.47**	80.09	19.43	79.79	18.84
	F1/RL	***81.64***	27.03	82.12	**33.98**	**83.33**	24.25	82.17	24.13
Llama-2-13b-chat									
	P/R1	84.84	***35.47***	84.21	***47.92***	84.94	37.87	85.89	23.33
	R/R2	79.14	20.63	79.96	29.73	79.95	19.35	78.48	12.01
	F1/RL	***81.88***	24.48	82.02	32.43	82.36	23.93	82.01	15.65
Llama-2-70b-chat									
	P/R1	84.39	***39.31***	84.26	35.48	85.33	***35.29***	85.40	25.64
	R/R2	78.34	25.09	79.25	17.61	79.94	18.30	78.41	12.56
	F1/RL	***81.24***	27.38	81.67	21.65	***82.54***	22.85	81.74	16.70
Mistral-7B-Instruct-v0.1									
	P/R1	85.29	***31.89***	85.61	36.64	86.86	30.59	86.26	29.47
	R/R2	79.14	20.81	79.38	25.77	80.09	18.49	79.74	17.98
	F1/RL	***82.08***	25.18	***82.37***	30.05	83.33	23.98	***82.85***	23.15
Mixtral-8x7B-Instruct-v0.1									
	P/R1	84.89	***35.45***	84.73	37.43	85.23	34.34	**85.28**	31.18
	R/R2	79.29	19.92	79.74	21.84	79.95	18.23	79.34	16.63
	F1/RL	***81.98***	25.39	**82.15**	27.07	82.49	24.39	***82.19***	22.06

TABLE IX. PERFORMANCE OF LLMS ON CNN/DM DATASET WITH ICL

Prompts		1-Shot		3-Shots		5-Shots		7-Shots	
Metrics		BERT	ROUGE	BERT	ROUGE	BERT	ROUGE	BERT	ROUGE
gemma-7b-it									
	P/R1	**84.04**	21.78	87.92	33.70	85.40	27.55	85.19	28.36
	R/R2	**82.27**	7.75	86.43	12.37	85.59	9.36	87.30	10.68
	F1/RL	**83.14**	12.60	87.16	22.06	85.47	17.96	86.22	18.55
Llama-2-7b-chat									
	P/R1	86.91	***35.92***	86.14	***32.58***	87.27	***34.35***	88.19	***39.45***
	R/R2	87.52	13.55	88.08	12.77	87.41	14.62	88.59	16.49
	F1/RL	**87.20**	23.01	87.09	**20.85**	***87.29***	23.16	88.37	26.25
Llama-2-13b-chat									
	P/R1	87.54	***36.31***	86.15	***32.37***	87.41	33.95	88.63	**39.43**
	R/R2	87.84	13.71	87.94	12.63	87.16	14.24	89.12	16.48
	F1/RL	**87.67**	23.11	87.02	20.70	87.24	22.62	88.86	26.22
Llama-2-70b-chat									
	P/R1	**87.63**	**38.26**	85.31	29.80	87.09	**33.11**	**88.97**	**40.98**
	R/R2	**87.68**	**14.51**	88.03	11.91	87.62	14.11	88.45	**17.23**
	F1/RL	**87.65**	**24.22**	86.63	19.23	**87.31**	22.32	**88.70**	**27.52**
Mistral-7B-Instruct-v0.1									
	P/R1	85.87	**32.47**	87.40	36.02	87.45	35.90	87.55	34.41
	R/R2	88.26	13.59	87.60	15.70	88.21	14.79	87.22	14.52
	F1/RL	**87.02**	21.19	**87.47**	23.80	87.81	23.70	**87.36**	23.26
Mixtral-8x7B-Instruct-v0.1									
	P/R1	86.75	**34.34**	**88.03**	**37.53**	**87.59**	**37.12**	**87.37**	36.60
	R/R2	86.31	12.74	**88.90**	**16.01**	**88.95**	**15.49**	**89.52**	15.74
	F1/RL	**86.51**	21.57	**88.44**	**24.79**	**88.24**	**24.58**	***88.41***	24.25

gains, particularly for the ROUGE and BERT metrics. These findings underscore the importance of demonstration-based prompting, reinforcing ICL's effectiveness in leveraging LLM capabilities for summarization.

B. Comparisons with State-of-the-Art Models

To situate LLMs within the broader landscape of text summarization, we compared their results with earlier state-of-the-art neural models across datasets Tables XIII and XIV. We focus on the highest-performing LLMs in our study in each dataset and contrast them with earlier existing models.

On news summarization (CNN/DM, NEWSROOM) in Table XIII, LLMs like Llama-2-70b-chat underperform models such as PEGASUS but eliminate dataset-specific training costs, as seen in Gemma-7b-it's NEWSROOM results.

For dialogue summarization, Table XIV shows that while Mixtral-8x7B demonstrates moderate performance, it slightly lags behind specialized models like SICK. This suggests that LLMs require further optimization for conversational coherence. On scientific documents, Table XIV shows that Llama-2-7b-chat achieves a high ROUGE-1 (49.74) but struggles with coherence (ROUGE-L: 33.98), reflecting sensitivity to prompts and context truncation.

These results highlight LLMs' potential for multi-domain adaptability but underscore the need for tailored prompts and hybrid architectures to balance semantic flexibility with syntactic precision.

C. Summarization with Chunking Strategy

To evaluate the impact of chunking, we tested it on the ArXiv dataset. Table XV reports ROUGE-1 and BERTScore F1 results before and after applying chunking.

Results show notable improvements for the Llama-2-70B-chat and Mixtral-8x7B-Instruct-v0.1 models in terms of both lexical and semantic metrics. However, models like Llama-2-13B-chat and gemma-7b-it did not consistently benefit, which

TABLE X. PERFORMANCE OF LLMS ON NEWSROOM DATASET WITH ICL

Examples		1 Shot		3 Shots		5 Shots		7 Shots	
Metric		**BERT**	**ROUGE**	**BERT**	**ROUGE**	**BERT**	**ROUGE**	**BERT**	**ROUGE**
gemma-7b									
	P/R1	84.4	13.1	86.4	20.3	85.4	18.0	80.7	10.7
	R/R2	84.1	3.0	86.6	5.8	87.1	5.6	86.7	3.7
	F1/RL	84.2	11.1	86.5	15.9	86.2	13.8	83.5	8.1
Llama2-7b									
	P/R1	82.4	*16.6*	85.1	21.6	**86.6**	20.9	86.3	**26.6**
	R/R2	85.5	5.0	88.2	7.1	85.8	6.2	88.4	**10.4**
	F1/RL	83.9	13.6	86.6	16.1	86.1	17.2	87.3	**21.2**
Llama2-13b									
	P/R1	82.3	*16.8*	84.7	***22.0***	83.6	23.0	86.9	26.3
	R/R2	85.5	5.0	87.8	7.2	84.8	9.3	88.8	10.0
	F1/RL	83.8	13.4	86.2	16.4	84.1	18.3	87.8	20.8
Llama2-70b									
	P/R1	82.7	17.3	85.3	20.9	84.5	22.6	87.3	26.0
	R/R2	85.8	5.2	88.1	6.6	86.3	7.7	88.5	9.2
	F1/RL	84.2	13.8	86.6	15.6	85.4	17.5	87.9	20.2
Mistral-7B									
	P/R1	**85.0**	19.1	85.8	23.4	86.2	23.6	87.2	22.8
	R/R2	**86.6**	6.6	87.4	8.3	87.8	8.8	86.0	8.4
	F1/RL	**85.8**	15.0	86.5	17.9	87.0	18.4	86.6	19.1
Mixtral-8x7B									
	P/R1	84.8	**20.6**	**86.3**	**23.7**	85.9	**24.3**	**84.0**	21.1
	R/R2	86.3	**6.2**	**88.6**	**8.0**	**88.4**	**9.2**	**87.5**	8.0
	F1/RL	85.5	**16.3**	**87.4**	**18.0**	**87.1**	**18.3**	***85.6***	16.3

TABLE XI. PERFORMANCE OF LLMS ON SAMSUM DATASET WITH ICL

		1 Shot		3 Shots		5 Shots		7 Shots	
		BERT	**ROUGE**	**BERT**	**ROUGE**	**BERT**	**ROUGE**	**BERT**	**ROUGE**
gemma-7b									
	P/R1	84.4	8.6	89.4	34.2	90.3	36.1	84.2	23.3
	R/R2	85.6	3.4	89.2	11.6	89.7	12.5	89.6	8.1
	F1/RL	85.0	6.7	89.3	27.1	90.0	28.8	86.8	17.4
Llama2-7b									
	P/R1	89.3	*39.1*	88.2	35.1	88.4	36.5	90.7	42.5
	R/R2	90.4	14.1	91.5	13.5	90.4	13.5	91.7	17.8
	F1/RL	89.9	31.0	89.8	27.0	89.4	28.2	91.2	34.2
Llama2-13b									
	P/R1	89.3	*39.1*	88.9	***39.4***	88.7	39.3	90.1	42.5
	R/R2	90.3	14.1	91.3	15.9	90.6	15.5	91.3	17.7
	F1/RL	89.8	31.1	90.1	30.8	89.6	30.5	90.6	34.0
Llama2-70b									
	P/R1	89.8	39.3	88.1	36.4	88.9	37.9	90.8	44.1
	R/R2	90.8	13.7	91.9	14.7	91.2	14.4	91.7	18.6
	F1/RL	90.3	31.0	89.9	28.2	90.0	29.5	91.2	35.7
Mistral-7B									
	P/R1	**89.5**	40.8	90.0	42.3	90.5	42.0	90.4	44.8
	R/R2	**91.6**	16.5	91.5	18.3	91.1	17.6	91.8	20.5
	F1/RL	**90.5**	32.2	90.7	34.0	90.8	33.7	91.0	36.0
Mixtral-8x7B									
	P/R1	**90.4**	**41.6**	**90.4**	**43.2**	**91.0**	**44.4**	**91.0**	**45.9**
	R/R2	**92.1**	**16.2**	**91.9**	**18.5**	**92.1**	**20.0**	**92.2**	**20.8**
	F1/RL	**91.2**	**32.6**	**91.1**	**34.6**	**91.6**	**35.8**	***91.6***	**37.0**

indicates that the chunking effectiveness may depend on the model architecture and input handling.

V. DISCUSSION

A. Prompt Diversity and In-Context Learning Effects

The datasets used in this study span distinct summarization domains—news (CNN/DM, NewsRoom), scientific documents (ArXiv), and dialogues (SAMSum)—each requiring tailored summarization strategies. To account for this variability, multiple prompts were designed and tested for each dataset to explore their influence on LLM behavior.

Fig. 2 illustrates how prompt variation affects ZSL performance. The results confirm that LLMs are highly sensitive to prompt phrasing: some prompts elicit more effective responses, improving both lexical and semantic quality, whereas others underperform. This highlights the critical role of prompt formulation in guiding LLMs across domains.

Our findings also demonstrate the adaptability of LLMs to different prompts and tasks. Their ability to generalize without fine-tuning highlights the potential of prompt engineering as a flexible tool. In addition, in-context learning (ICL) significantly boosts performance by providing demonstration pairs, especially as the number of examples increases. This reinforces the value of contextual learning in terms of improving model comprehension and summary generation quality.

TABLE XII. PERFORMANCE OF LLMS ON ARXIV DATASET WITH ICL

Prompts	1 Shot		3 Shots		5 Shots		7 Shots	
Metrics	BERT	ROUGE	BERT	ROUGE	BERT	ROUGE	BERT	ROUGE
gemma-7b								
P/R1	81.5	22.4	83.7	30.1	85.4	35.2	83.2	29.6
R/R2	78.5	2.6	81.5	7.5	82.3	10.7	81.2	7.1
F1/RL	80.0	13.5	82.6	17.2	83.8	20.6	82.2	17.7
Llama2-7b								
P/R1	85.7	40.8	86.7	42.2	86.1	48.3	86.9	39.8
R/R2	82.1	13.7	**84.2**	15.0	**83.9**	17.8	84.0	14.3
F1/RL	**83.8**	22.6	**85.4**	24.0	85.0	22.5	85.4	23.3
Llama2-13b								
P/R1	**85.8**	40.7	**86.7**	**42.1**	**86.7**	47.9	87.0	40.0
R/R2	81.8	13.7	84.1	**15.1**	83.6	19.4	83.9	14.5
F1/RL	83.7	22.7	85.4	**24.0**	**85.1**	22.1	85.4	23.4
Llama2-70b								
P/R1	85.9	**41.0**	86.7	41.7	85.7	39.4	86.5	**40.4**
R/R2	81.9	**13.7**	82.8	14.7	82.7	15.7	83.5	**14.1**
F1/RL	83.8	**22.8**	84.7	23.8	84.2	18.6	85.0	**23.0**
Mistral-7B								
P/R1	**84.0**	37.1	84.8	37.5	85.6	38.9	85.2	37.0
R/R2	**82.1**	12.8	82.8	12.7	82.1	12.9	81.6	12.7
F1/RL	**83.1**	20.0	83.8	20.6	83.8	21.5	83.3	20.6
Mixtral-8x7B								
P/R1	84.5	39.3	85.7	40.4	86.6	40.1	85.9	39.5
R/R2	80.9	13.1	83.6	14.0	83.5	13.9	84.4	13.4
F1/RL	82.6	21.8	84.6	22.3	85.0	22.9	*85.1*	21.8

TABLE XIII. COMPARISON OF TRADITIONAL SUMMARIZATION MODELS AND LLMS ACROSS CNN/DM AND SAMSUM DATASETS

Dataset	Method	R-1	R-2	R-L
CNN/DM	Hie-BART[31]	44.35	21.37	41.05
	PEGASUS[30]	44.17	21.47	41.1
	Llama-2-70b-chat	40.98	17.23	27.52
NEWSROOM	PEGASUS[30]	45.15	33.51	41.33
	Gemma-7b-it	25.38	8.52	19.96

TABLE XIV. COMPARISON OF TRADITIONAL SUMMARIZATION MODELS AND LLMS ON NEWSROOM AND ARXIV DATASETS

Dataset	Method	R-1	R-2	R-L
SAMSum	SICK[11]	53.73	28.81	49.5
	Mixtral-8x7B	45.86	20.78	36.96
ArXiv	PRIMERA[7]	47.60	20.80	42.60
	PEGASUS[30]	44.70	17.27	25.80
	Llama-2-7b-chat	49.74	32.47	33.98

TABLE XV. PERFORMANCE COMPARISON BEFORE/AFTER CHUNKING ON ARXIV DATASET (ROUGE-1 & BERTSCORE F1)

Model	ROUGE-1		BERTScore	
	Before	After	Before	After
Gemma-7B	**40.05**	39.91	**84.51**	83.60
Llama2-7B	49.74	**50.30**	83.92	**84.06**
Llama2-13B	**47.92**	47.01	**84.21**	82.04
Llama2-70B	35.48	**38.12**	84.26	**86.19**
Mistral-7B	36.64	**36.90**	85.61	**85.80**
Mixtral-8x7B	37.43	**40.02**	84.73	**87.16**

B. LLMs in Summarizing Long Documents

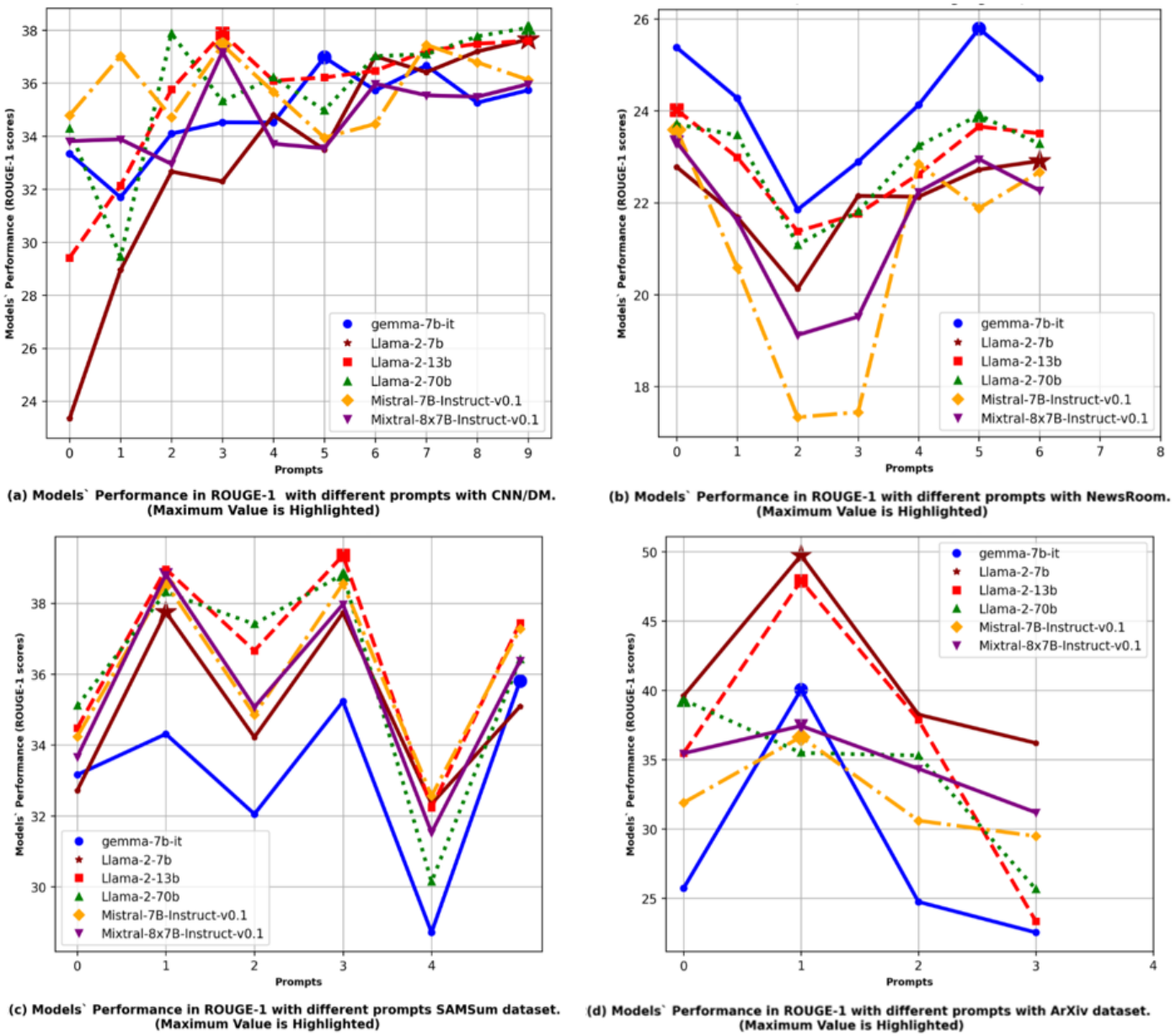
As shown in Table XV, chunking improved summarization performance—particularly ROUGE-1 and BERTScore—for models like Llama-2-70B-chat and Mixtral-8x7B-Instruct-v0.1, indicating better preservation of contextual and semantic content. In contrast, models such as Mistral-7B-Instruct-v0.1, Llama-2-13B-chat, andgemma-7b-it showed minimal or inconsistent gains, which suggests that the chunking effectiveness varies according to the model architecture and context handling capabilities.

While chunking helps retain key information in lengthy inputs and improves final summary quality, it introduces additional inference overhead. This trade-off between performance and efficiency is particularly important when processing long-form scientific texts like those in the ArXiv dataset.

C. Comparison of Generated Summary Lengths with Reference Summaries

Another analysis is conducted regarding the results summaries generated from the employed models to gain more insight into the results. Hence, the lengths of the generated summaries were compared to those of the reference summaries for all datasets used in this study. Table XVI offers a summary of the average lengths of the summaries produced by each model when given different prompts, together with the reference summary lengths for each dataset. The average length was calculated by taking the average lengths of the resulting summaries across the different prompts used. The average lengths of the generated summaries demonstrate that the LLMs do not consistently generate summaries that are comparable to the reference summaries. The results demonstrate that most models generate summaries that are longer than the reference summaries.

Overall, these findings imply that the strengths and weaknesses of various models vary according to the dataset characteristics, in addition to model's length, number of parameters, and prompt building, as previously described. For instance, Llama-2-13b-chat excels in producing summaries that are nearly the reference length for datasets like SAMSum and ArXiv, and gemma-7b-it consistently generates succinct summaries for most datasets.



(a) Models` Performance in ROUGE-1 with different prompts with CNN/DM. (Maximum Value is Highlighted)

(b) Models` Performance in ROUGE-1 with different prompts with NewsRoom. (Maximum Value is Highlighted)

(c) Models` Performance in ROUGE-1 with different prompts SAMSum dataset. (Maximum Value is Highlighted)

(d) Models` Performance in ROUGE-1 with different prompts with ArXiv dataset. (Maximum Value is Highlighted)

Fig. 2. Change in models' performance in ROUGE-1 across datasets using multiple prompts in zero-shot prompting.

However Mistral-7B-Instruct-v0.1 shows a tendency to produce longer summaries that may risk verbosity in an attempt to capture more detail. Consequently, the characteristics of the particular dataset and the intended ratio of detail to conciseness should be taken into account when choosing a model for a summarization task. In conclusion, there is a trade-off between verbosity and the efficiency of the generated summaries when using LLMs for ATS, and this trade-off affects how well these summaries are evaluated in comparison to reference summaries.

D. Variation in LLM Performance: Insights and Contributing Factors

Performance variations across models and datasets reflect the complexity of summarization tasks. Larger models—such as Llama-2-70B-chat and Mixtral-8x7B—tend to perform better, especially on complex datasets like CNN/DM and ArXiv; however, this trend is not universal [53], [54].

Prompt design significantly affects performance. Here, structured, summary-specific prompts generally yield more relevant and coherent outputs than open-ended queries. ICL further enhances model understanding by providing context-rich demonstrations, and performance improves as the number of examples increases [41].

However, no single factor alone accounts for performance differences. Instead, the interplay among model size, architecture, prompt formulation, and demonstration count shapes summarization quality. Continued investigation is needed to isolate and optimize these factors for improved LLM perfor-

TABLE XVI. AVERAGE GENERATED SUMMARY LENGTHS OF LLMS COMPARED TO REFERENCE SUMMARY LENGTHS

	Datasets			
Models	**CNN/DM**	**NewsRoom**	**SAMSum**	**ArXiv**
gemma-7b-it	**53.584**	**27.672**	44.722	143.31
Llama-2-7b-chat	83.456	56.8	47.534	289.3575
Llama-2-13b-chat	65.706	47.67	**39.308**	**208.825**
Llama-2-70b-chat	71.864	44.832	40.45	200.4975
Mistral-7B-Instruct-v0.1	108.056	60.46	48.098	168.57
Mixtral-8x7B-Instruct-v0.1	83.704	45.07	46.528	204.71
Reference Length (#Words)	**52**	**26**	**22**	**220**

mance on abstractive summarization tasks.

E. Qualitative Analysis of Generated Summaries

We further analyzed the summaries of the top-performing CNNDM, NewsRoom, and SAMSum models. Key qualitative observations are:

1) Content coverage: Despite lexical differences, summaries often capture core meanings and key facts. Many outputs paraphrase reference summaries while retaining relevance.

2) Coherence: Most outputs maintain logical flow and clarity, even with varied stylistic choices, reflecting LLMs' capacity for fluent text generation.

3) Error patterns: Common issues include omissions, factual distortions, and overgeneralizations (e.g., "more than 200 people" simplified as "a large crowd"). NewsRoom outputs sometimes misstate facts; CNNDM examples show reduced specificity.

In conclusion, LLM-generated summaries generally preserve the intended meaning and demonstrate strong contextual relevance, even when they diverge from human-written references in phrasing. This illustrates their potential to produce effective abstractive summaries.

F. Inference Time Analysis

We also measured the average inference time required by each model to generate summaries across the datasets. Table XVII shows these results.

TABLE XVII. AVERAGE INFERENCE TIME (SECONDS) PER ARTICLE ACROSS MODELS AND DATASETS

Model	CNN/DM	NewsRoom	ArXiv	SAMSum
Llama2-7b-chat	23.34	25.1	61.2	11.5
Llama2-13b-chat	25.7	30.3	68.8	14.9
Llama2-70b-chat	80.6	95.2	148.4	30.7
Gemma-7b-it	21.1	23.8	63.4	10.2
Mistral-7b-instruct	17.9	20.5	52.9	8.3
Mixtral-8×7b-instruct-v0.1	31.5	36.8	100.3	15.4

Three key patterns emerged: (1) Larger models incur higher latency—e.g., Llama2-70B-chat significantly outpaces its smaller variants; (2) Inference time scales with input length—ArXiv documents take longer to process than shorter SAMSum dialogues; and (3) Architectural optimization matters—Mixtral-8x7B, despite its size, is faster than Llama2-70B on ArXiv due to its sparse MoE design.

These findings emphasize the trade-off between performance and efficiency and highlight the need to balance quality and latency in practical deployments.

VI. CONCLUSION

This study presents a comprehensive evaluation of locally hosted LLMs on news, dialogue, and scientific text summarization tasks using zero-shot learning (ZSL), in-context learning (ICL), and chunking strategies. Our findings highlight several key insights. First, LLM performance varies significantly across domains and models. While models like Mixtral-8x7B-Instruct and Llama-2-70B-chat demonstrated strong results in dialogue and scientific summarization, smaller models such as Gemma-7b-it struggled with long-form inputs. Second, ICL consistently outperformed ZSL in terms of providing contextual examples, which resulted in improved lexical and semantic alignment. Third, the chunking strategy proved beneficial for handling long scientific documents, particularly enhancing outputs from models constrained by limited context windows. Additionally, qualitative analysis of the generated summaries revealed that many LLMs effectively preserved the intended meaning and maintained contextual coherence even when their outputs diverged lexically from human references. Common error patterns—such as omissions, generalizations, and factual distortions—were identified, providing further guidance for model refinement and deployment. Performance differences were influenced by factors such as models architectural considerations, scale, prompt tuning, and the number of in-context examples. Overall, this work underscores the promise of open-source LLMs as adaptable tools for abstractive summarization and offers both quantitative and qualitative insights into their effective use across varied textual domains. Future research could examine optimizing LLMs for domain-specific summarization, particularly for long-form scientific texts where chunking introduces coherence challenges. In addition, exploring retrieval-augmented and integrating external knowledge sources may improve factuality and context awareness. Optimizing the inference efficiency in real-time or low-resource environments will continue to be a crucial area for practical deployment considerations.

REFERENCES

[1] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Systems with Applications*, vol. 165, p. 113679, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420305030

[2] A. Tamkin, M. Brundage, J. Clark, and D. Ganguli, "Understanding the capabilities, limitations, and societal impact of large language models," 2021.

[3] A. Abdallah and A. Jatowt, "Generator-retriever-generator: A novel approach to open-domain question answering," *arXiv preprint arXiv:2307.11278*, 2023.

[4] V. Dehru, P. K. Tiwari, G. Aggarwal, B. Joshi, and P. Kartik, "Text summarization techniques and applications," *IOP Conference Series: Materials Science and Engineering*, vol. 1099, no. 1, p. 012042, mar 2021. [Online]. Available: https://dx.doi.org/10.1088/1757-899X/1099/1/012042

[5] J. He, W. Kryscinski, B. McCann, N. Rajani, and C. Xiong, "Ctrlsum: Towards generic controllable text summarization," *ArXiv*, vol. abs/2012.04281, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:227745074

[6] P. Gigioli, N. Sagar, A. Rao, and J. Voyles, "Domain-aware abstractive text summarization for medical documents," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2018, pp. 2338–2343.

[7] W. Xiao, I. Beltagy, G. Carenini, and A. Cohan, "PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5245–5263. [Online]. Available: https://aclanthology.org/2022.acl-long.360/

[8] S. Rafi and R. Das, "Topic-guided abstractive multimodal summarization with multimodal output," *Neural Computing and Applications*, 2023. [Online]. Available: https://doi.org/10.1007/s00521-023-08821-5

[9] N. Moratanch and S. Chitrakala, "A survey on extractive text summarization," in *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*, 2017, pp. 1–6.

[10] S. Chitrakala and N. Moratanch, "A survey on abstractive text summarization," in *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, 2016, pp. 1–7.

[11] S. Kim, S. J. Joo, H. Chae, C. Kim, S.-w. Hwang, and J. Yeo, "Mind the gap! injecting commonsense knowledge for abstractive dialogue summarization," in *Proceedings of the 29th International Conference on Computational Linguistics*, N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, Eds. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 6285–6300. [Online]. Available: https://aclanthology.org/2022.coling-1.548/

[12] A. Dingare, D. Bein, W. Bein, and A. Verma, "Abstractive text summarization using machine learning," in *ITNG 2022 19th International Conference on Information Technology-New Generations*, S. Latifi, Ed. Cham: Springer International Publishing, 2022, pp. 269–276.

[13] P. Kouris, G. Alexandridis, and A. Stafylopatis, "Abstractive text summarization based on deep learning and semantic content generalization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 5082–5092. [Online]. Available: https://aclanthology.org/P19-1501

[14] Y. Li, Y. Huang, W. Huang, J. Yu, and Z. Huang, "An abstractive summarization model based on joint-attention mechanism and a priori knowledge," *Applied Sciences*, vol. 13, no. 7, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/7/4610

[15] J. Pilault, R. Li, S. Subramanian, and C. Pal, "On extractive and abstractive neural document summarization with transformer language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 9308–9319. [Online]. Available: https://aclanthology.org/2020.emnlp-main.748

[16] J. Li, C. Zhang, X. Chen, Y. Cao, P. Liao, and P. Zhang, "Abstractive text summarization with multi-head attention," in *2019 International Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8.

[17] Y. Liu, K. Shi, K. S. He, L. Ye, A. R. Fabbri, P. Liu, D. Radev, and A. Cohan, "On learning to summarize with large language models as references," 2023.

[18] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, jan 2023. [Online]. Available: https://doi.org/10.1145/3560815

[19] B. Chen, Z. Zhang, N. Langren'e, and S. Zhu, "Unleashing the potential of prompt engineering in large language models: a comprehensive review," *ArXiv*, vol. abs/2310.14735, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264426395

[20] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: Techniques and applications," 2024. [Online]. Available: https://arxiv.org/abs/2402.07927

[21] X. Amatriain, "Prompt design and engineering: Introduction and advanced methods," *ArXiv*, vol. abs/2401.14423, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267301483

[22] "The inverted pyramid - purdue owl - purdue university," accessed: 2024-01-22.

[23] R. Nallapati, F. Zhai, and B. Zhou, "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents," 2016.

[24] Q. Zhou, N. Yang, F. Wei, S. Huang, M. Zhou, and T. Zhao, "Neural document summarization by jointly learning to score and select sentences," 2018.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[27] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3730–3740. [Online]. Available: https://aclanthology.org/D19-1387

[28] V. G. M. Gambhir, "Deep learning-based extractive text summarization with word-level attention mechanism," *Multimed Tools Appl*, vol. 81, p. 20829–20852, 2022.

[29] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. rahman Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Annual Meeting of the Association for Computational Linguistics*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204960716

[30] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization," 2020.

[31] K. Akiyama, A. Tamura, and T. Ninomiya, "Hie-BART: Document summarization with hierarchical BART," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, E. Durmus, V. Gupta, N. Liu, N. Peng, and Y. Su, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 159–165. [Online]. Available: https://aclanthology.org/2021.naacl-srw.20

[32] L. Basyal and M. Sanghvi, "Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models," 2023.

[33] B. I. L Tang, Z Sun, "Evaluating large language models on medical evidence summarization," *npj Digit. Med*, vol. 6, pp. –, 2023.

[34] D. V. Veen, C. V. Uden, L. Blankemeier, J.-B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E. P. Reis, A. Seehofnerova, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C. P. Langlotz, J. Hom, S. Gatidis, J. Pauly, and A. S. Chaudhari, "Clinical text summarization: Adapting large language models can outperform human experts," 2023.

[35] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, X. Sun, L. Li, and Z. Sui, "A survey on in-context learning," 2024. [Online]. Available: https://arxiv.org/abs/2301.00234

[36] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gulçehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, S. Riezler and Y. Goldberg, Eds. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 280–290. [Online]. Available: https://aclanthology.org/K16-1028

[37] M. Grusky, M. Naaman, and Y. Artzi, "Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 708–719. [Online]. Available: http://aclweb.org/anthology/N18-1065

[38] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, "Samsum corpus: A human-annotated dialogue dataset for abstractive summarization," in *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Association for Computational Linguistics, 2019. [Online]. Available: http://dx.doi.org/10.18653/v1/D19-5409

[39] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, "A discourse-aware attention model for abstractive summarization of long documents," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 615–621. [Online]. Available: https://aclanthology.org/N18-2097

[40] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, jan 2019. [Online]. Available: https://doi.org/10.1145/3293318

[41] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?" in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 11 048–11 064. [Online]. Available: https://aclanthology.org/2022.emnlp-main.759

[42] X. He, G. Haffari, and M. Norouzi, "Sequence to sequence mixture model for diverse machine translation," in *Proceedings of the 22nd Conference on Computational Natural Language Learning*, A. Korhonen and I. Titov, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 583–592. [Online]. Available: https://aclanthology.org/K18-1056

[43] M. Dalal, A. C. Li, and R. Taori, "Autoregressive models: What are they good for?" *ArXiv*, vol. abs/1910.07737, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204743691

[44] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[45] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," *ArXiv*, vol. abs/2310.06825, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263830494

[46] "Gemma," https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf, accessed: 2024-02-5.

[47] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de Las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mixtral of experts," *ArXiv*, vol. abs/2401.04088, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:266844877

[48] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, "A survey on mixture of experts," 2024. [Online]. Available: https://arxiv.org/abs/2407.06204

[49] H. Gong, Y. Shen, D. Yu, J. Chen, and D. Yu, "Recurrent chunking mechanisms for long-text machine reading comprehension," 2020. [Online]. Available: https://arxiv.org/abs/2005.08056

[50] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

[51] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," 2020.

[52] M. Peeperkorn, T. Kouwenhoven, D. G. Brown, and A. K. Jordanous, "Is temperature the creativity parameter of large language models?" *ArXiv*, vol. abs/2405.00492, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269484653

[53] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.

[54] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, and M. M. Kabir, "A survey of automatic text summarization: Progress, process and challenges," *IEEE Access*, vol. 9, pp. 156 043–156 070, 2021.