Domain Adaptation for Multi-document Summarisation: A Case Study in the Medical Research Domain

Kushan Hewapathirana[1,2] Nisansa de Silva[1] C.D. Athuraliya[2] Piumi Kandanaarachchi[3]
[1]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
[2]ConscientAI, Sri Lanka
[3]District General Hospital, Hambantota, Sri Lanka
{kushan.22, nisansa}@cse.mrt.ac.lk
cd@conscient.ai, piumikandanaarachchi@gmail.com

Abstract

Effectively summarising medical research is critical for supporting evidence-based decision-making in healthcare. While fine-tuning task-specific models on domain data is established practice, the comparative advantages over increasingly capable general-purpose LLMs remain an open question. This study systematically evaluates domain-adapted PRIMERA against several open-source large language models (LLaMA 3.2 3B, Mistral 7B, OpenChat 7B, and Gemma 7B) in zero-shot settings using the MS^2 dataset, which includes $20,000$ systematic reviews summarising over $470,000$ medical studies. Fine-tuning leads to notable improvements in ROUGE scores—ROUGE-1 from 12.8 to 33.0, ROUGE-2 from 2.0 to 6.5, and ROUGE-L from 8.1 to 22.6. Comparative evaluation indicates that the fine-tuned model consistently achieves stronger performance across all three ROUGE metrics, human evaluations, and LLM-as-a-judge assessments. These results suggest that domain-adapted models can offer advantages over general-purpose LLMs in specialised settings, particularly where factual accuracy and coverage are critical, though at the cost of reduced flexibility across domains.

1 Introduction

Multi-document summarisation (MDS) is a challenging Natural Language Processing (NLP) task that aims to generate a summary by combining information from multiple sources. MDS involves handling conflicting, duplicate or complementary information to produce a summary that represents the overall content (Hewapathirana et al., 2023). The goal of MDS is to condense a collection of documents into a single, cohesive summary that captures the main points and ideas of the original documents (Ma et al., 2023; Afsharizadeh et al., 2022; Abid, 2022). Automatic summarisation can be classified into two primary categories: extractive and abstractive. *Extractive text summaries* contain keywords, phrases, and sentences that are extracted verbatim from the source documents (Ma et al., 2023; Afsharizadeh et al., 2022; Pasunuru et al., 2021), whereas *abstractive text summaries* generate summaries that include paraphrased sentences and new terms that may not be found in the original documents (Ma et al., 2023; Afsharizadeh et al., 2022; Pasunuru et al., 2021; Abid, 2022).

MDS can involve summarising different types of documents, including short sources, long sources, and hybrid sources. Short sources are documents such as tweets, product reviews, or headlines that convey a smaller amount of information. In contrast, long sources are lengthy documents such as news articles or research papers that contain a large amount of information and detail. Hybrid sources contain one or few long documents with several to many short documents, such as a scientific summary from a long paper with several corresponding citations (Ma et al., 2023; Afsharizadeh et al., 2022; Pasunuru et al., 2021; Yu, 2022; Abid, 2022; Hewapathirana et al., 2023; Wolhandler et al., 2022).

MDS researchers use various techniques to generate abstractive summaries, such as natural language generation, deep learning models, and neural machine translation. These techniques enable the automatic creation of summaries that are coherent, informative, and useful (Ma et al., 2023; Afsharizadeh et al., 2022; Hewapathirana et al., 2024).

This study addresses three key research questions: (1) How does domain-adapted fine-tuning of a task-specific MDS model compare with task-specific baselines and recent open-source LLMs in zero-shot settings? (2) To what extent do fine-tuned models generalise beyond their training dataset, across domains and document types? (3) How do automatic metrics (ROUGE (Lin, 2004)) align with human evaluation and LLM-as-a-judge assessments in the medical summarisation domain?

Our main contributions include: A compara-

tive evaluation of fine-tuned PRIMERA against task-specific models (PEGASUS, LED) and open-source LLMs (LLaMA 3.2, Mistral, Gemma, OpenChat); empirical evidence demonstrating that domain adaptation provides measurable advantages over zero-shot approaches in specialized settings where factual accuracy is critical; and an analysis of generalization capabilities across document types, revealing both the strengths and limitations of domain-specific fine-tuning.

2 Related Work

MDS has evolved significantly with the introduction of transformer-based architectures. Early models such as BigBird (Zaheer et al., 2020) and Longformer (Beltagy et al., 2020) addressed the challenge of handling long input sequences via sparse and sliding window attention. These architectures were extended in summarisation tasks by replacing standard self-attention mechanisms in models like BART (Lewis et al., 2020), enabling more efficient long-context encoding. Hierarchical encoders have also been explored to capture inter-document structure more effectively.

Pre-trained transformer models such as BERTSUM (Liu and Lapata, 2019), BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020), and T5 (Raffel et al., 2020) have established strong baselines for abstractive summarisation. These models leverage large-scale pre-training to capture rich contextual information and have demonstrated high-quality generation across a variety of tasks. PRIMERA (Xiao et al., 2022), a Longformer Encoder-Decoder (LED)-based model trained with an entity-based pyramid pretraining strategy, has shown state-of-the-art performance in MDS benchmarks.

Domain-specific approaches such as CGSUM (Chen et al., 2022), which uses citation-guided selection for summarising scientific papers, and DAMEN (Moro et al., 2022), which incorporates indexing and discriminative filtering for medical MDS, illustrate the benefit of incorporating structural or domain-aware features into summarisation pipelines.

Recent work has also examined the challenge of synthesising sentiment or conflicting perspectives across multiple documents. DeYoung et al. (2024) proposed using Diverse Beam Search (Vijayakumar et al., 2016) to generate a range of candidate summaries, selecting the one most representative of the aggregate view. This improves robustness to variations in input structure and composition.

For training and evaluation, benchmark datasets such as DUC and TAC [1] [2] have historically been used, although they suffer from size and positional bias. More recent alternatives include Multi-News (Fabbri et al., 2019), WikiSum (Liu et al., 2018), and WikiHow (Koupaee and Wang, 2018), which offer larger and more diverse summary corpora. Additionally, datasets such as Rotten Tomatoes (Leon, 2020) have supported the evaluation of aggregation quality across subjective inputs.

In parallel, the rise of large language models (LLMs) has significantly impacted summarisation. While proprietary models such as GPT-4 and Claude have shown strong results, their closed nature and resource demands limit reproducibility (Laskar et al., 2023). Open-access models such as LLaMA 3.2 3B (Meta AI, 2024), Mistral 7B (Jiang et al., 2023), Gemma 7B (Team et al., 2024), and OpenChat 7B (Wang et al., 2023) offer competitive performance in summarisation while remaining lightweight enough for deployment in constrained academic or clinical environments. These models enable researchers to explore LLM-based summarisation in low-resource settings without sacrificing modern capabilities.

3 Model Selection

In this study, we selected models based on a combination of empirical performance in MDS, architectural diversity, domain relevance, and computational feasibility. Our selection process was guided by a thorough review of recent literature and benchmarking studies, with ROUGE scores (Lin, 2004) and adoption in the MDS community serving as key criteria.

After careful evaluation, we chose to assess the performance of three summarisation models: PRIMERA (Xiao et al., 2022), PEGASUS (Zhang et al., 2020), and Longformer Encoder-Decoder (LED) (Beltagy et al., 2020). PRIMERA is a state-of-the-art MDS model that leverages the Longformer architecture and an entity pyramid masking strategy to enhance content selection during pre-training. It has consistently outperformed earlier methods in benchmark evaluations (Afsharizadeh et al., 2022; Ma et al., 2023; DeYoung et al., 2024). PEGASUS, with its Gap Sentence Gen-

[1] https://duc.nist.gov/
[2] https://tac.nist.gov/

eration (GSG) objective, is particularly effective in generating summary-worthy sentences and has demonstrated strong performance on abstractive summarisation tasks. LED, with its sparse attention mechanism, serves as a strong baseline due to its efficiency in handling long input sequences and its use in earlier MDS studies.

In addition to these task-specific models, we evaluated a set of recent open-access LLMs with strong general-purpose summarisation capabilities: LLaMA 3.2 3B (Meta AI, 2024), Mistral 7B (Jiang et al., 2023), Gemma 7B (Team et al., 2024), and OpenChat 7B (Wang et al., 2023). These models represent lightweight alternatives to proprietary commercial LLMs and are increasingly being adopted in low-resource and open research settings. Though not explicitly fine-tuned for MDS, their instruction-following abilities allow for effective few-shot or zero-shot summarisation, making them valuable for comparative evaluation against domain-specific models.

These LLMs were selected based on availability, performance in recent summarisation benchmarks, and their suitability for deployment under academic resource constraints. All experiments involving LLMs were conducted on a server with a 4-core CPU, 64 GB RAM, and a single A2 GPU with 16 GB VRAM, which limited the feasibility of larger commercial models and motivated the use of accessible open-weight alternatives.

Together, these model selections enable a comprehensive comparison across specialised, pretrained summarisation models and general-purpose LLMs in the medical MDS setting.

3.1 MS^2 Dataset

Existing summarisation datasets often lack biomedical specificity, limiting their effectiveness for domain-specific summarisation tasks (Ma et al., 2023; Afsharizadeh et al., 2022; Abid, 2022). To address this, DeYoung et al. (2021) introduced the MS^2 dataset, specifically curated for biomedical document summarisation using systematic literature reviews. These reviews synthesize evidence from multiple studies, providing concise and clinically relevant summaries. e.g., a review on Vitamin B12 supplementation in older adults may aggregate diverse findings (Andrès et al., 2010).

The dataset was constructed by filtering the Semantic Scholar Open Research Corpus (Lo et al., 2020) using a multi-stage pipeline: keyword heuristics to identify systematic reviews (220K), a PubMed filter for biomedical relevance (170K), and a SciBERT-based classifier (Beltagy et al., 2019) for final selection, yielding 20K high-quality review-summary pairs. This makes MS^2 a robust benchmark for training and evaluating biomedical summarisation models with strong relevance to evidence-based healthcare applications.

Dataset		PRIMERA	PEGASUS	LED
Multi-News	R-1	**42.0***	32.0*	17.3*
	R-2	**13.6***	10.1*	3.7*
	R-L	**20.8***	16.7*	10.4*
Multi-Xscience	R-1	**29.1***	27.6*	14.6*
	R-2	**4.6***	4.6*	1.9*
	R-L	**15.7***	15.3*	9.9*
WikiSum	R-1	**28.0***	24.6*	10.5*
	R-2	**8.0***	5.5*	2.4*
	R-L	**18.0***	15.0*	8.6*
BigSurvey-MDS	R-1	23.9$^{\diamond}$	**38.9**†	39.8†
	R-2	4.1$^{\diamond}$	**9.0**†	9.4†
	R-L	11.7$^{\diamond}$	**16.2**†	16.1†
MS^2	R-1	12.8$^{\diamond}$	12.7$^{\diamond}$	**25.8**‡
	R-2	2.0$^{\diamond}$	1.5$^{\diamond}$	**8.4**‡
	R-L	8.1$^{\diamond}$	8.3$^{\diamond}$	**19.3**‡
Rotten Tomatoes	R-1	25.4$^{\bullet}$	**27.4**$^{\bullet}$	25.6$^{\bullet}$
	R-2	8.4$^{\bullet}$	**9.5**$^{\bullet}$	8.0$^{\bullet}$
	R-L	19.8$^{\bullet}$	**21.1**$^{\bullet}$	19.6$^{\bullet}$

Table 1: **ROUGE scores of selected models across different domains.** Datasets include *Multi-News* (Fabbri et al., 2019), *Multi-Xscience* (Lu et al., 2020), *WikiSum* (Liu et al., 2018), *BigSurvey-MDS* (Liu et al., 2023), *MS^2* (DeYoung et al., 2021), and *Rotten Tomatoes* (Leon, 2020). Sources: * Xiao et al. (2022), † Liu et al. (2023), ‡ DeYoung et al. (2021), $^{\bullet}$ Wang et al. (2022), $^{\diamond}$ Hewapathirana et al. (2023).

3.2 Evaluation Metrics

We evaluated the fine-tuned PRIMERA model using a combination of automated metrics, human evaluation, and LLM-based judgments to comprehensively assess its performance on the MS^2 dataset.

Automated Metrics. The primary automated metric used was ROUGE (Lin, 2004), a standard for MDS evaluation (Ma et al., 2023). ROUGE measures the overlap between the generated summary and the reference summary, focusing on precision and recall. We employed two key variants: ROUGE-N, which calculates *n*-gram overlap, and ROUGE-L, which assesses sentence-level similarity based on the longest common subsequence (Lin, 2004). These metrics enabled objective comparison across baseline and state-of-the-art models.

Human Evaluation. To assess qualitative performance, we conducted human evaluations focusing on five criteria: *Relevance*, *Coherence*, *Coverage*, *Conciseness*, and *Accuracy*. Three expert annotators from the medical domain independently rated 50 randomly sampled summaries. Inter-annotator agreement was measured using Krippendorff's Alpha (α) (Krippendorff, 1989), which is robust to multiple raters and missing data. This human assessment provided critical insight into the linguistic and domain-specific fidelity of the summaries.

LLM-as-a-Judge Evaluation To further assess summary quality from a model-based perspective, we employed the DeepEval framework[3] using Meta's LLaMA 3 90B Instruct model (`us.meta.llama3-2-90b-instruct-v1:0`) hosted via AWS Bedrock. Evaluation was conducted on a sample of 50 summaries due to cost constraints. By adopting DeepEval's standardized summarisation evaluation protocol[4], we ensured reproducibility and methodological consistency. This LLM-based evaluation complemented ROUGE and human assessments by providing judgments on factual consistency, coherence, and overall quality.

Domain-Specific Fine-Tuning. The medical domain represents a unique challenge due to its complexity and specialized vocabulary. To address this, we fine-tuned PRIMERA using the MS^2 dataset, comprising medical research papers. Our objective was to improve PRIMERA's performance in generating accurate and concise summaries tailored for biomedical literature.

Training Configuration. Fine-tuning was performed with carefully chosen hyperparameters: a learning rate of $5e$-05, batch size of 4, and 3 training epochs. We used the Adam optimizer with betas (0.9, 0.999) and epsilon $1e$-08, along with a linear learning rate scheduler. A random seed of 42 ensured reproducibility. The Hugging Face `trainer` API[5] was used to manage the training process efficiently on our available infrastructure.

[3] https://github.com/confident-ai/deepeval

[4] https://deepeval.com/docs/metrics-summarization

[5] https://huggingface.co/docs/transformers/main_classes/trainer

4 Results

4.1 Fine-tuned Model Performance

The fine-tuning of the PRIMERA model on the MS^2 dataset resulted in significant improvements in performance. Prior to fine-tuning, the ROUGE-1, ROUGE-2, and ROUGE-L scores were observed to be approximately 12.8, 2.0, and 8.1, as shown in Table 1. However, after fine-tuning the model, these scores significantly increased to 33.0, 6.5, and 22.6, as presented in Table 2. Notably, the fine-tuned PRIMERA model outperformed the state-of-the-art LED model which achieved ROUGE-1 and ROUGE-L scores of 25.8 and 19.3 respectively.

Table 2: Performance of our fine-tuned PRIMERA model on various domain-specific datasets

Dataset		PRIMERA
Multi-News	R-1	39.1
	R-2	11.7
	R-L	18.0
BigSurvey-MDS	R-1	33.0
	R-2	7.1
	R-L	12.7
MS^2	R-1	33.0
	R-2	6.5
	R-L	22.6

In order to assess the generalization capabilities of the fine-tuned model, we also evaluated its performance on a different domain dataset, the Multi-news dataset (Fabbri et al., 2019). This dataset primarily consists of news articles and their corresponding human-written summaries from the website `newser.com`. It encompasses a diverse range of news sources, making it more representative of real-world scenarios compared to previous datasets such as DUC and Newsroom (Fabbri et al., 2019). The fine-tuned model exhibited slightly lower ROUGE scores when tested on the Multi-news dataset. Although there was a slight drop in performance compared to the initial PRIMERA model performances, the results remained reasonable. Furthermore, we evaluated the fine-tuned model on the BigSurvey dataset (Liu et al., 2023), which consists of survey papers and their corresponding summaries. The dataset includes two levels of target summaries: a comprehensive long summary and a concise short summary. The fine-tuned PRIMERA model demonstrated improved performance on the BigSurvey dataset as well.

These findings demonstrate mixed generalization patterns. While the fine-tuned model shows strong performance on MS^2 and maintains im-

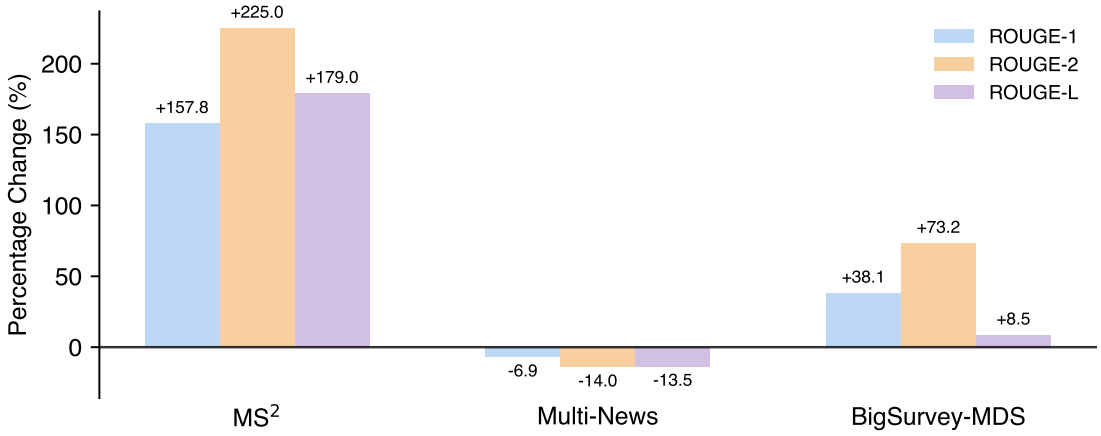

Figure 1: Percentage Improvement/Reduction of Fine-Tuned PRIMERA Model

provements on the structurally similar BigSurvey dataset (+38.1% R-1, +73.2% R-2, +8.5% R-L), it experiences modest performance degradation on Multi-News (−6.9% R-1, −14.0% R-2, −13.5% R-L). This suggests the model has adapted to the specific characteristics of medical research papers rather than learning a fully generalizable MDS strategy. The improved performance on BigSurvey likely reflects similarity in document structure (both are research papers with systematic organization) rather than pure domain transfer. These results indicate that domain-specific fine-tuning trades broad generalization for targeted performance gains, which may be acceptable or even desirable when the deployment context matches the training domain.

A comparison on the percentage change on all three datasets is given in Figure 1.

4.2 Zero-Shot Evaluation of Open LLMs

To understand how well recent open-source LLMs perform in the medical domain without task-specific fine-tuning, we evaluated several zero-shot models on the MS^2 dataset. Specifically, we selected four high-performing and resource-accessible models: LLaMA 3.2 3B, Mistral 7B, Gemma 7B, and OpenChat 7B. These models were chosen based on their availability, instruction-following capabilities, and competitive performance in prior evaluations on general summarisation tasks.

The evaluation was performed using the ROUGE metric suite to maintain consistency with our fine-tuned PRIMERA results. Table 3 presents the ROUGE-1, ROUGE-2, and ROUGE-L F1 scores obtained for each model.

These results demonstrate that while open LLMs exhibit a basic ability to generate summaries, their performance lags behind the fine-tuned PRIMERA model on this domain-specific dataset. For instance, the best-performing zero-shot model (LLaMA 3.2 3B) achieves only 18.73 in ROUGE-1, compared to 33.0 for fine-tuned PRIMERA. Notably, all zero-shot models struggled with ROUGE-2, indicating limited ability to capture fine-grained bi-gram-level details common in dense, technical summaries.

Table 3: Zero-shot ROUGE performance of open LLMs on MS^2.

Model	ROUGE-1	ROUGE-2	ROUGE-L
LLaMA 3.2 3B	18.73	3.07	10.97
Mistral 7B	18.40	3.56	11.80
OpenChat 3.5	17.80	3.43	11.20
Gemma 7B	15.60	2.90	9.85

These findings highlight the value of domain adaptation in specialised tasks such as medical summarisation, where zero-shot summarisation remains limited by lack of domain-specific knowledge and training alignment.

4.3 Qualitative Analysis in the Context of Medical Domain

In addition to showcasing the effectiveness of fine-tuned models for medical research summarisation, it is important to highlight the significant benefits that this model brings to the field of healthcare and medical research. Medical research papers are often dense and lengthy, containing a wealth of information that is crucial for healthcare professionals, researchers, and policymakers. However, the sheer volume and complexity of these papers can make it challenging to extract the key findings and insights efficiently, resulting in either a poor analysis of a

research paper or complete disregard of the publication due to the difficulty in summarising and coming to a conclusion on its findings. Moreover, focusing solely on the abstract and conclusion for the sake of efficiency may cause the reader to overlook important concepts as well. Hence, having a system in place to efficiently review the entire paper and generate a summary in a fraction of the time would be highly beneficial to any medical personnel.

Generated

The results of this meta-analysis suggest that surgical ablation of the left atrial fibrillation (LAF) is associated with a significantly lower incidence of recurrent atrial arrhythmias compared with catheter ablation. However, there was no significant difference in the incidence of major adverse events between the two groups. Surgical ablation appears to be a safe and effective treatment strategy for the treatment of LAF.

Ground Truth

Subgroup analysis demonstrated similar trends, with higher freedom from AF in the surgical ablation group for paroxysmal AF patients. The incidence of pacemaker implantation was higher, while no difference in stroke or cardiac tamponade was demonstrated for the surgical versus catheter ablation groups. Current evidence suggests that epicardial ablative strategies are associated with higher freedom from AF, higher pacemaker implantation rates and comparable neurological complications and cardiac tamponade incidence to catheter ablative treatment.

Figure 2: Comparison of Generated Summary from Our Model and Ground Truth.

Figure 2 showcases a comparison between the generated summary produced by our fine-tuned PRIMERA model and the corresponding ground truth summary. To aid the visualization, we have distinguished sentences in both summaries using different colors to demonstrate similarities and differences between them. To ensure a reliable evaluation of the summary's quality, considering the challenging task of human evaluation for all generated summaries, we conducted a sample analysis on a small subset of generated summaries and ground truth pairs using domain experts. In the figure, we have highlighted certain sentences in different colors to represent specific findings. Yellow - Express an agreement that surgical ablation yields better results than catheter ablation for AF patients, Green - Indicate an agreement that the occurrence of adverse events does not significantly differ between the two groups, namely surgical ablation and catheter ablation groups, Blue - Facts that are different between the two statements. The first is the usage of the term "LAF" which is medically inaccurate, and the second is epicardial ablation being mentioned in the ground truth in place of surgical ablation but entirely missing from the generated summary.

An analysis of the summaries generated by the proposed model revealed both advantages and disadvantages when compared to the ground truth. The generated statement fully aligned with two out of the three points of the ground truth, with only a few minor errors. Nevertheless, the word "epicardial ablation" has been used in place of "surgical ablation" in the ground truth, which is conceptually correct and would be easily understood by a healthcare professional who is familiar with the subject. However, deciphering this has been difficult for the algorithm given the complexity of the content. On the other hand, the term "Left Atrial Fibrillation" or "LAF" is used which is medically incorrect with regards to this research. Therefore it is evident that developing a flawless system for summarising complex content remains a significant challenge.

To further examine how the fine-tuned model compares to recent open LLMs, we performed a qualitative comparison using the LLaMA 3.2 3B model as an illustrative example. Although the LLaMA model generated syntactically fluent outputs, it frequently omitted critical findings and occasionally introduced factual inconsistencies. For instance, as shown in Table 4, summaries from LLaMA often presented overly generic conclusions or contradicted key results stated in the source documents.

As shown in Table 4, PRIMERA outputs closely aligned with the factual content and structure of the ground truth summaries, capturing nuanced medical relationships such as treatment hierarchies and combined efficacy. In contrast, LLaMA summaries either ignored critical qualifiers or introduced misleading claims; such as denying evidence where it actually exists. These issues are particularly concerning in clinical contexts where accuracy is non-negotiable.

This comparison highlights that while open LLMs can offer a strong baseline in zero-shot setups, they fall short in medical domain-specific summarisation without adaptation. Despite the resource-intensive nature of LLMs, our findings demonstrate that smaller, fine-tuned models like PRIMERA can deliver superior factual accuracy and domain relevance. These qualitative differences further support the value of targeted domain

Table 4: Sample summaries comparing ground truth, PRIMERA, and LLaMA 3.2 3B outputs.

Ground Truth	PRIMERA Output	LLaMA 3.2 3B Output
PDE5Is were significantly more effective than placebo or SSRIs for treating PE, while SSRIs were better than placebo. Combined treatment had better efficacy than PDE5Is alone.	PDE5Is were more effective than SSRIs or placebo. Combined treatments were more effective than PDE5Is alone.	There is no evidence to support the use of SSRIs for treating PE. Insufficient data to determine their effectiveness.
Silver-level evidence concerning the beneficial effects of mineral baths compared to no treatment. No clear effects for other balneological treatments were found.	Mineral baths showed beneficial effects over no treatment. Effects of other treatments were unclear.	Balneotherapy is effective for osteoarthritis of the knee in adults.
Oral cobalamin improves serum vitamin B12 and hematological parameters. Avoids discomfort and cost of injections. Supported for clinical use.	Oral cobalamin is effective in improving serum B12 and blood parameters and avoids injections.	Oral cobalamin is effective for vitamin B12 deficiency, but there is limited evidence supporting either oral or intramuscular use.

adaptation to ensure safe and reliable summarisation in high-stakes fields like healthcare.

4.4 Inter-annotator Agreement

To evaluate the subjectivity and consistency among human annotators, we used Krippendorff's Alpha (α) (Krippendorff, 1989), a robust statistical measure for inter-rater reliability that is well-suited to ordinal-scale annotations and missing data scenarios. Unlike Cohen's Kappa (Cohen, 1960), Krippendorff's Alpha supports multiple annotators and provides a more generalizable reliability estimate. Cohen's Kappa, while commonly used, assumes only two raters and is less robust to missing data, making Krippendorff's Alpha a more appropriate choice for our setting. Three medical experts independently evaluated 50 randomly selected samples—both ground truth and generated summaries—against the original documents. Each summary was scored on five key criteria: Relevance, Coherence, Coverage, Conciseness, and Accuracy—using a rubric co-developed with a medical expert (See Appendix A). The results, shown in Table 5, reveal consistently moderate inter-annotator agreement, with generated summaries demonstrating slightly higher agreement across all criteria.

Notably, criteria such as Coverage ($\alpha = 0.5541$) and Coherence ($\alpha = 0.5038$) exhibited the highest reliability for generated outputs, indicating that the model produces content that aligns more consistently with expert expectations. Interestingly, annotator agreement was higher for generated summaries than for ground truth in all five evaluation dimensions. This suggests that, while subjectivity remains inherent to human judgment, the generated summaries (PRIMERA) offer a more stable and interpretable baseline, especially when paired with a domain-aligned rubric.

Evaluation Criterion	Ground Truth	Generated Summary
Relevance	0.3012	0.4512
Coherence	0.4531	0.5038
Coverage	0.4215	0.5541
Conciseness	0.2536	0.3534
Accuracy	0.3824	0.4817

Table 5: Krippendorff's Alpha scores for ground truth and generated summaries across five evaluation criteria.

While Krippendorff's Alpha provides a robust measure of agreement, the relatively small sample size (sample of 50 summaries, compared to the full dataset of $20,000$ reviews) limits the strength of claims we can make about population-level properties. The observed patterns suggest interesting trends—particularly the higher agreement on generated summaries—but should be interpreted as preliminary findings that warrant validation on larger samples in future work. Statistical significance testing was not performed due to the small sample size and the ordinal nature of the data.

4.5 LLM-as-a-Judge Results

As introduced in the Evaluation Metrics section, we employed DeepEval's LLM-as-a-judge framework to compare 50 summaries generated by the fine-tuned PRIMERA model and four open-source LLMs—LLaMA 3.2 3B, Mistral 7B, OpenChat 7B, and Gemma 7B—using Meta's LLaMA 3 90B Instruct model (`us.meta.llama3-2-90b-instruct-v1:0`) as the judge.

The comparative G-Eval scores for all models are summarised in Table 6.

The results indicate that PRIMERA, a domain-adapted model, consistently outperformed the

Table 6: Scores for PRIMERA and Open LLMs (Sample size: 50)

Metric	PRIMERA	LLaMA 3.2 3B	Mistral 7B	OpenChat 7B	Gemma 7B
Avg. Score	0.3452	0.2732	0.2850	0.2780	0.2650
Std. Deviation	0.1274	0.0788	0.0821	0.0773	0.0755
Median Score	0.3461	0.2655	0.2750	0.2700	0.2600
Max Score	0.6273	0.4434	0.4650	0.4520	0.4310
Min Score	0.1254	0.1309	0.1420	0.1355	0.1282

evaluated open-source LLMs across all metrics. Among the zero-shot baselines, Mistral 7B achieved the highest G-Eval scores, followed by OpenChat 7B, LLaMA 3.2 3B, and Gemma 7B.

Although the open LLMs demonstrated general summarisation capability in zero-shot settings, their performance declined on criteria such as factual consistency and coverage when applied to medical texts. These results suggest that, in specialised domains such as healthcare, domain-specific fine-tuning remains necessary to achieve reliable and contextually accurate summarisation.

5 Discussion

This study provides a systematic comparison of domain-adapted task-specific models versus general-purpose LLMs for medical MDS. While fine-tuning on domain-specific data is now standard practice, and improved performance on that domain is expected, the key question we addressed is whether such adaptation offers meaningful advantages over increasingly capable zero-shot LLMs, particularly in resource-constrained settings where deploying large models may be impractical.

Our comparative evaluation reveals that domain adaptation continues to provide substantial benefits for specialized summarisation tasks. The fine-tuned PRIMERA model consistently outperformed all evaluated open-source LLMs (LLaMA 3.2 3B, Mistral 7B, OpenChat 7B, Gemma 7B) across automated metrics, human judgments, and LLM-as-a-judge assessments. Critically, this performance advantage came with significantly lower computational requirements—PRIMERA can run efficiently on modest hardware, while larger LLMs demand substantial resources even for inference.

To further examine generalisation after fine-tuning, we evaluated the model on out-of-domain datasets. On Multi-News, which predominantly consists of news articles, performance declined modestly relative to the pre-fine-tuned PRIMERA baseline (-6.9% to -14.0% across ROUGE metrics), consistent with mild catastrophic forgetting. By contrast, on the BigSurvey dataset of survey papers, the fine-tuned model showed clear gains ($+8.5\%$ to $+73.2\%$), indicating stronger adaptability within research-style domains that share structural similarity with biomedical papers.

Beyond ROUGE-based performance trends, we examined qualitative aspects of summarisation through human evaluation. Expert ratings offered complementary insights into coherence, factual accuracy, and domain relevance, providing a human-centred view of fine-tuning effects. Summaries from the fine-tuned PRIMERA model were rated higher across relevance, coherence, coverage, conciseness, and accuracy, and were generally more structured and focused than some human-written references, though minor factual inconsistencies remained. These findings motivated further reliability analysis using inter-annotator agreement and automated overlap metrics.

In this study, we evaluated the quality of machine-generated summaries for MDS in the medical domain using two key methodologies: inter-annotator agreement analysis via Krippendorff's Alpha (Krippendorff, 1989) and content overlap comparison using ROUGE metrics. Krippendorff's Alpha (α), chosen for its robustness to multiple raters and missing data, provided a more reliable estimate of human agreement than traditional methods like Cohen's Kappa. The aim was to assess both the consistency of expert annotations and the relative quality of generated summaries compared to human-written ones.

To complement traditional metrics, we also compared PRIMERA against several zero-shot open LLMs using LLM-as-a-judge evaluations. Despite reasonable performance by open models (e.g., Mistral in G-Eval), PRIMERA consistently outperformed them across all metrics. As shown in Table 6, the G-Eval scores further reinforce these findings. PRIMERA attained the highest average score (0.3452), outperforming all zero-shot baselines, with Mistral 7B (0.2850) and LLaMA 3.2 3B (0.2732) following. Scores closer to 1 indicate stronger factual consistency and coherence; thus, the ≈ 0.07 - 0.08 margin highlights a clear qualitative advantage. PRIMERA also showed slightly higher variance (SD = 0.1274), reflecting diverse yet consistently strong outputs.

This reinforces the conclusion that fine-tuned domain-specific models continue to offer critical advantages, especially in specialised, high-stakes domains such as biomedical summarisation. While

open LLMs offer flexibility and broad applicability, their outputs often lack the factual consistency and specificity demanded by domain-expert tasks. The qualitative error analysis also supported this finding—highlighting key factual inaccuracies and omissions in LLM-generated outputs.

Although large open LLMs are powerful, their resource requirements for inference and deployment can be prohibitive in many real-world settings. Pre-trained models like PRIMERA, when fine-tuned on task-specific datasets such as MS^2, demonstrate competitive performance with significantly lower computational demands. This trade-off highlights a practical benefit of domain-adapted summarisation models: they offer a scalable, cost-effective alternative while still maintaining high-quality performance.

Therefore, while larger open models may continue to improve, our findings suggest that fine-tuned pre-trained models remain a highly valuable and robust solution for domain-specific summarisation, particularly in scenarios where resource efficiency and reliability are paramount.

6 Conclusion

This study explored the domain adaptation of the PRIMERA model for MDS, with a focus on the biomedical domain using the MS^2 dataset. Through systematic fine-tuning, we demonstrated that the adapted model significantly outperforms the pre-trained baseline and competitive models such as LED, particularly in ROUGE metrics. Furthermore, our evaluations across multiple domains, including news and survey articles, indicate that the fine-tuned PRIMERA model retains strong generalisation capabilities beyond its training domain.

To provide a more nuanced perspective on summary quality, we incorporated LLM-as-a-judge evaluations using the DeepEval framework. These results further confirmed that the fine-tuned PRIMERA model surpasses leading open-source LLMs (e.g., LLaMA 3.2 3B, Mistral 7B) in terms of factual accuracy, coherence, and overall summarisation quality. While open LLMs showed promise in zero-shot settings, they underperformed in capturing domain-specific nuances critical for biomedical content.

Qualitative analyses echoed this gap, highlighting factual inconsistencies in zero-shot summaries, whereas PRIMERA more reliably retained core evidence and reasoning. Importantly, the resource efficiency of PRIMERA—compared to computationally intensive LLMs—positions it as a practical solution for real-world deployment in specialised domains.

Our findings demonstrate that domain-adapted task-specific models remain valuable for specialized summarisation, particularly when factual accuracy, computational efficiency, and deployment constraints are considerations. While the performance advantages of fine-tuning on domain data are expected, this study provides empirical evidence that such adaptation offers meaningful benefits over zero-shot LLMs in medical contexts—though at the cost of reduced generalization across domains.

Limitations

While this study provides convergent evidence across automated, human, and LLM-based evaluations, several limitations should be acknowledged. First, both the human and LLM-as-a-judge assessments were conducted on a small qualitative subset of 50 summaries (approximately 0.25% of the test set) due to computational constraints. Although this sample offers indicative comparative trends, it limits the statistical strength of our conclusions and precludes significance testing.

Second, open-source LLMs were evaluated only in zero-shot settings. Fine-tuning these models on the MS^2 dataset or employing few-shot prompting could substantially improve their performance, potentially narrowing the observed performance gap with PRIMERA.

Third, our evaluation was restricted to the biomedical domain. While cross-domain tests on news and survey papers provided preliminary evidence of generalisation, further investigation across other specialised domains—such as legal, financial, and technical documentation—is needed to assess the broader applicability of domain adaptation.

Future work should therefore incorporate larger-scale human and LLM-based evaluations, compare fine-tuned general-purpose LLMs against domain-specific models, and explore domain-aware accuracy metrics tailored for medical summarisation. Ultimately, the choice between fine-tuned task-specific models and general-purpose LLMs will depend on deployment context, balancing computational resources, domain sensitivity, and generalisation needs.

References

Azal Minshed Abid. 2022. Multi-Document Text Summarization Using Deep Belief Network.

Mahsa Afsharizadeh, Hossein Ebrahimpour-Komleh, Ayoub Bagheri, and Grzegorz Chrupała. 2022. A Survey on Multi-document Summarization and Domain-Oriented Approaches. *Journal of Information Systems and Telecommunication (JIST)*, 1(37):68.

Emmanuel Andrès, Helen Fothergill, and Mustapha Mecili. 2010. Efficacy of oral cobalamin (vitamin b12) therapy. *Expert opinion on pharmacotherapy*, 11(2):249–256.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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)*, pages 3615–3620. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150*.

Jingqiang Chen, Chaoxiang Cai, Xiaorui Jiang, and Kejia Chen. 2022. Comparative graph-based summarization of scientific papers guided by comparative citations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5978–5988. International Committee on Computational Linguistics.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and psychological measurement*, 20(1):37–46.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MS^2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513. Association for Computational Linguistics.

Jay DeYoung, Stephanie C. Martinez, Iain J. Marshall, and Byron C. Wallace. 2024. Do multi-document summarization models synthesize? *Transactions of the Association for Computational Linguistics*, 12:1043–1062.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084. Association for Computational Linguistics.

Kushan Hewapathirana, Nisansa De Silva, and C D Athuraliya. 2023. Multi-Document Summarization: A Comparative Evaluation. In *2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS)*, pages 19–24. IEEE.

Kushan Hewapathirana, Nisansa de Silva, and C D Athuraliya. 2024. M2DS: Multilingual Dataset for Multi-document Summarisation. In *International Conference on Computational Collective Intelligence*, pages 219–231. Springer.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Mahnaz Koupaee and William Yang Wang. 2018. WikiHow: A Large Scale Text Summarization Dataset. *arXiv preprint arXiv:1810.09305*.

Klaus Krippendorff. 1989. *Content analysis: an introduction to its methodology*. Sage Publications.

Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. 2023. Building real-world meeting summarization systems using large language models: A practical perspective. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 343–352. Association for Computational Linguistics.

Stefano Leon. 2020. Rotten tomatoes movies and critic reviews dataset. `https://bit.ly/RTdataset`. (Accessed on 06/24/2023).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. GENERATING WIKIPEDIA BY SUMMARIZING LONG SEQUENCES. In *International Conference on Learning Representations*.

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2023. Generating a Structured Summary of Numerous Academic Papers: Dataset and Method. In *THE 31ST INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE*, pages 4259–4265.

Yang Liu and Mirella Lapata. 2019. 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), pages 3730–3740. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983. Association for Computational Linguistics.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074. Association for Computational Linguistics.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2023. Multi-document Summarization via Deep Learning Techniques: A Survey. *ACM Computing Surveys*, 55(5):1–37.

Meta AI. 2024. Llama 3.2 model card. https://huggingface.co/meta-llama/Llama-3.2-1B. Accessed: 2025-08-16.

Gianluca Moro, Luca Ragazzi, Lorenzo Valgimigli, and Davide Freddi. 2022. Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 180–189. Association for Computational Linguistics.

Ramakanth Pasunuru, Mengwen Liu, Mohit Bansal, Sujith Ravi, and Markus Dreyer. 2021. Efficiently summarizing text and graph encodings of multi-document clusters. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4768–4779. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of machine learning research*, 21(140):1–67.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 19 others. 2024. Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. DIVERSE BEAM SEARCH: DECODING DIVERSE SOLUTIONS FROM NEURAL SEQUENCE MODELS. *arXiv preprint arXiv:1610.02424*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Lucy Lu Wang, Jay DeYoung, and Byron Wallace. 2022. Overview of MSLR2022: A shared task on multi-document summarization for literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 175–180. Association for Computational Linguistics.

Ruben Wolhandler, Arie Cattan, Ori Ernst, and Ido Dagan. 2022. How "multi" is multi-document summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5761–5769. Association for Computational Linguistics.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. 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)*, pages 5245–5263. Association for Computational Linguistics.

Benjamin Yu. 2022. Evaluating pre-trained language models on multi-document summarization for literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 188–192. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for Longer Sequences. *Advances in neural information processing systems*, 33:17283–17297.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

A Human Evaluation Rubric for Summary Assessment

This rubric was used by medical domain experts to evaluate both ground truth and model-generated summaries. Each summary was assessed independently across five criteria using a five-point Likert scale.

A.1 Evaluation Scale

Criterion-Specific Guidelines

Relevance Does the summary capture key clinical findings?

Score	Interpretation
1	Very Poor — fails completely in this aspect
2	Poor — significant issues are present
3	Fair — partially meets expectations, with flaws
4	Good — mostly meets expectations, minor issues
5	Excellent — fully meets expectations

Table 7: Likert scale used across all evaluation criteria.

1: Largely irrelevant or misleading.

2: Omits several key points.

3: Covers some relevant points; others diluted.

4: Includes most important findings.

5: Captures all clinically essential points.

Coherence Is the summary logically structured and easy to follow?

1: Confusing or disjointed.

2: Poor logical flow.

3: Basic structure with inconsistencies.

4: Mostly well-structured, minor issues.

5: Clear, fluent, and logically organized.

Coverage Does the summary reflect the breadth of the source?

1: Misses most critical content.

2: Narrow focus; limited scope.

3: Some key elements included.

4: Broadly representative.

5: Fully covers main findings.

Conciseness Is the summary free from redundancy or unnecessary detail?

1: Excessively verbose.

2: Frequent redundancy.

3: Some inefficiencies.

4: Mostly succinct.

5: Highly concise and focused.

Accuracy Are the statements factually correct?

1: Contains major factual errors.

2: Several inaccuracies.

3: Some vague or incorrect content.

4: Mostly accurate, minor issues.

5: Fully accurate and aligned with source.

Instructions to Annotators

- Read all source documents before scoring a summary.
- Evaluate each criterion independently.
- Use the rubric definitions to ensure consistency.
- When uncertain, assign the most justifiable score.