DomainSum: A Hierarchical Benchmark for Fine-Grained Domain Shift in Abstractive Text Summarization

Haohan Yuan **Haopeng Zhang**[*]
ALOHA Lab, University of Hawaii at Manoa
{haohany,haopengz}@hawaii.edu

Abstract

Most research on abstractive summarization focuses on single-domain applications, often neglecting how domain shifts between documents affect performance and the generalization ability of summarization models. To address this issue, we introduce DomainSum, a hierarchical benchmark designed to capture fine-grained domain shifts in abstractive summarization. We categorize these shifts into three levels: genre, style, and topic, and demonstrate through comprehensive benchmark analysis that they follow a hierarchical structure. Furthermore, we evaluate the domain generalization capabilities of commonly used pre-trained language models (PLMs) and large language models (LLMs) in in-domain and cross-domain settings. Our benchmark and source code are released at `https://github.com/hpzhang94/DomainSum`.

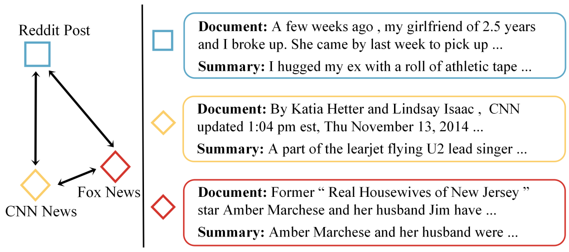


Figure 1: An illustration of varying domain shifts between content styles, with smaller shifts between news articles (CNN vs. Fox) and larger shifts between news articles and Reddit posts.

1 Introduction

Abstractive summarization is a crucial task in natural language processing (NLP) that aims to generate concise and coherent summaries by interpreting and distilling essential information from a source text. As the volume of publicly available text continues to grow exponentially, the demand for automatic summarization methods has become more urgent. Recent advancements in pre-trained language models (PLMs) and large language models (LLMs) have significantly improved the performance of abstractive summarization systems, achieving unprecedented results in generating human-like summaries (Liu and Lapata, 2019; Liu et al., 2022; Zhang et al., 2020a, 2023d).

However, much of the current research focuses on summarizing specific types of documents, such as news articles or academic papers (Zhang et al., 2023c, 2024a). This narrow focus limits the models' ability to generalize across documents with diverse characteristics, hindering their effectiveness in real-world applications. Specifically, many models struggle to adapt to different content types due to unaddressed distributional discrepancies in the training data, a phenomenon commonly referred to as summarization domain shift (Wang et al., 2019).

Recently, researchers have begun investigating how domain-specific corpus characteristics affect summarization performance. Early studies (Hua and Wang, 2017) examined out-of-domain training for summarization models or utilized document categories and latent topics for multi-task learning (Cao et al., 2017). Wang et al. (2019) modeled summarization domain shift by analyzing distributional differences between news sources (e.g., CNN vs. New York Times), applying meta-learning and domain tagging to tackle multi-domain extractive summarization. Building on this, Yu et al. (2021) explored domain adaptation for abstractive summarization, emphasizing the distributional disparities across document types (e.g., emails vs. news vs. dialogues) and proposed continued pre-training for low-resource settings. More recently, Afzal et al. (2024) evaluated large language models (LLMs) for domain adaptation in zero-shot and few-shot settings, focusing on distributional differences across topics (e.g., science vs. medicine vs. government).

Despite these efforts, a clear and consistent def-

[*]corresponding author

Findings of the Association for Computational Linguistics: NAACL 2025, pages 2219–2231

inition of domain shift in summarization remains elusive, and current methods lack a fine-grained measurement to capture the variability of domain shifts across different content types. As illustrated in Fig 1, the domain shift between two news articles on different topics should be smaller than the shift between a news article and a Reddit post. However, existing studies and benchmarks fail to capture this nuanced distinction in domain shifts across varying types of content.

To address these limitations, we introduce DomainSum, a hierarchical benchmark designed to investigate fine-grained domain shifts in abstractive summarization. Inspired by the systematic formulation of language style (DiMarco and Hirst, 1993), we categorize domain shift into three distinct levels of granularity: genre shift, style shift, and topic shift, as depicted in Fig 2. We validate this categorization through a comprehensive corpus analysis, examining key properties such as compression ratio, density, and abstractiveness across these three levels. DomainSum is constructed from high-quality public datasets and encompasses five distinct domains for each level, creating a large-scale, hierarchical testbed. Furthermore, we evaluate the performance of both current PLMs and LLMs on our benchmark, uncovering their ability to handle varying degrees of domain shift. The key contributions of this paper are threefold:

- We categorize summarization domain shift into three levels (genre, style, and topic) and present DomainSum, a large-scale hierarchical benchmark that enables a comprehensive evaluation of summarization performance across diverse content types.
- We perform a detailed analysis of domain shift within DomainSum across different granular levels, examining eight domain characteristic measures such as compression, density, coverage, diversity, and abstractiveness across domains.
- We evaluate the domain shift capabilities of existing PLMs and LLMs, highlighting their performance variations across different granular levels of domain shift.

2 Related Work

2.1 Abstractive Summarization

Abstractive summarization aims to create concise and coherent summaries from scratch, often employing sequence-to-sequence models (Sutskever, 2014). In contrast to extractive summarization methods (Zhang et al., 2022, 2023a,b), abstractive summarization offers greater flexibility and fluency while reducing redundancy. The advent of PLMs has significantly advanced the field, leading to notable improvements in fluency, coherence, and informativeness (Lewis, 2019; Zhang et al., 2020a; Liu et al., 2022).

Recently, LLMs have gained considerable attention for summarization due to their strong in-context learning (ICL) capabilities (Brown, 2020) and chain-of-thought reasoning abilities (Wei et al., 2022; Wang et al., 2023). With minimal or no examples, LLMs can generate summaries that rival those produced by fine-tuned PLMs (Min et al., 2022; Afzal et al., 2024). For example, Goyal et al. (2022) found that while GPT-3 summaries yielded slightly lower ROUGE scores, human evaluators preferred them. Similarly, Zhang et al. (2024b) reported that LLM-generated summaries in the news domain performed comparably to those written by humans. Furthermore, Zhang et al. (2023d) introduced an iterative framework for text summarization, allowing LLMs to refine their outputs through self-evaluation and feedback, thereby improving both faithfulness and control.

2.2 Domain Shift for Summarization

Domain shift has been extensively studied in NLP and computer vision (CV) (Ganin et al., 2016; Gururangan et al., 2020; Ramponi and Plank, 2020), focusing on the distributional differences between training and test data. However, exploration of domain adaptation in abstractive summarization remains limited. Hua and Wang (2017) first investigated the adaptation of neural summarization models to out-of-domain data. Following this, Wang et al. (2019) examined domain shifts in news sources specifically for extractive summarization. Magooda and Litman (2020) expanded this area by introducing cross-domain data synthesis methods for summarization. Yu et al. (2021) addressed the adaptation of abstractive summarization across domains through continued pre-training. More recently, Afzal et al. (2024) explored the application of LLMs for zero-shot and few-shot adaptation across various topics. Additionally, Li et al. (2024) found that the learning difficulty of datasets exhibits an almost linear relationship with cross-domain overlap and performance gain.

Despite these advancements, there is still a lack

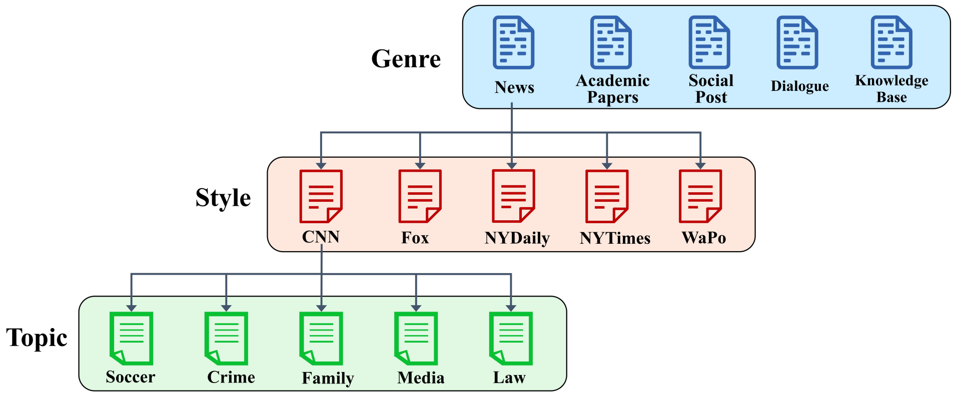


Figure 2: The overall hierarchical structure of DomainSum, featuring three granular levels of shifts (genre, style, and topic) and five distinct domains at each level.

of clear and consistent definitions of domain shift in summarization, as well as a significant gap in studying the nuanced influences of document distributional differences on summarization.

3 DomainSum Benchmark

This section introduces our DomainSum benchmark in detail. In Section 3.1, we outline the benchmark's construction process, which leverages high-quality public datasets. Additionally, we present a comprehensive analysis of DomainSum using eight summarization domain characteristic measures in Section 3.2.

3.1 Benchmark Construction

Domain refers to a metadata attribute that divides data into distinct parts with varying distributions (Joshi et al., 2012). Domain shift is crucial for supervised systems, as a mismatch between training and testing domains can adversely affect performance. In the context of abstractive summarization, a conditional generation task, domain shift captures the distributional differences between document-summary pairs. Specifically, we propose to categorize domain shifts in summarization into three granular levels: **genre shift, style shift, and topic shift**, drawing inspiration from the systematic formulation of language style (DiMarco and Hirst, 1993).

To capture nuanced distinctions in domain shifts across different types of content, we construct our DomainSum benchmark hierarchically. Specifically, the benchmark consists of three levels, each containing document-summary pairs from five distinct datasets. Notably, this structure is not a fully branched tree. The five datasets at the second level are exclusively derived from the first dataset at the preceding level. Similarly, the five datasets at the third level originate from the first dataset at the second level. The hierarchy is illustrated in Fig 2, and the details of the included data are as follows:

Genre Shift We begin by defining genre shift, which refers to the changes in content, format, and structure that arise when summarizing texts across different genres, such as news articles, academic papers, and fiction. This phenomenon presents distinct challenges in the summarization process, requiring models to adapt to the diverse conventions and audience expectations inherent to each genre.

For DomainSum, we leverage high-quality, publicly available datasets from five specific genres: **i) News:** We pick the widely-adopted CNN/DailyMail (Hermann et al., 2015) news summarization dataset, which contains news articles alongside human-written highlights as summaries. **ii) Academic papers:** We use the Arxiv dataset (Cohan et al., 2018) as a representative for scientific papers and long-form documents. Following the setup in (Zhong et al., 2020), the introduction section is treated as the document and the abstract as the summary. **iii) Social post:** We include the Reddit dataset (Kim et al., 2018), which contains highly abstractive summaries from social media. We specifically use the TIFU-long version, where the body text of a post is regarded as the document and the TL;DR as the summary. **iv) Dialogue:** The SAMSum dataset (Gliwa et al., 2019) is incorporated to represent text message-like dialogues. It comprises single-document conversations created by linguists fluent in English, covering a broad range of formality levels and topics.

Datasets	Train	Validation	Test
	Genre Shift		
CNN/DM	287,084	13,367	11,489
PubMed	83,233	4,946	5,025
Reddit	41,675	645	645
SAMSum	14,732	818	819
WikiHow	168,126	6,000	6,000
	Style Shift		
CNN	43,466	4,563	4,619
Fox	78,760	8,423	8,392
NYDaily	55,653	6,057	5,904
NYTimes	152,959	16,488	16,620
WaPo	95,379	9,939	10,072
	Topic Shift		
Soccer	16,313	1,889	1,546
Crime	24,597	1,131	973
Family	26,019	1,176	1,047
Media	24,364	1,189	919
Law	15,564	550	532

Table 1: Overview of the train, validation, and test splits for DomainSum.

v) Knowledge base: We include the WikiHow dataset (Koupaee and Wang, 2018), which consists of diverse instructional contents extracted from an online knowledge base. Each summary succinctly encapsulates the main steps or advice provided in the guide article.

Style Shift As one finer-grained step, even within the same genre, different sources or authors produce articles with varying styles, distinguished by differences in tone, length, word frequency, and polarization. We define this disparity between documents and their summaries as style shift.

To capture this variation, we focus on the news article genre and construct fine-grained domain shift data that reflects diverse styles of news articles. Following the settings in (Wang et al., 2019), we reuse and sample from the Newsroom corpus (Grusky et al., 2018), which comprises article-summary pairs authored by writers and editors from 38 major publications between 1998 and 2017. We selected the top five publications (The New York Times (**NYTimes**), The Washington Post (**WaPo**), Fox News (**Fox**), the New York Daily News (**NYDaily**), and CNN (**CNN**)) and processed the data in standardized formats. We specifically retained CNN news from the Newsroom dataset to ensure consistency in summaries across both genre and topic levels.

Topic Shift As a further step, documents within the same genre and style can still exhibit significant variation in terms of topics and emphasis. For instance, news articles may cover subjects such as sports or finance, incorporating different terminologies and catering to distinct audience expectations and summary information density.

To account for this variation, we closely examine the news genre and the aforementioned CNN news style, constructing a finer-grained domain shift that captures diverse topics within the news article-summary pairs. Specifically, we employ the Latent Dirichlet Allocation (LDA) topic model (Blei et al., 2003) to cluster the CNN news data by topic. From this analysis, we identify and include the five most frequent categories: **Soccer, Crime, Family, Media, and Law**. Additionally, we conduct human verification to manually filter out data with low topic relevance.

Altogether, the data instances from the three levels of summarization in DomainSum are organized in a hierarchical structure, capturing varying levels of granularity. Detailed statistics for the DomainSum benchmark, including the training, validation, and testing sets, are presented in Table 1.

3.2 Benchmark Analysis

Following the three granular levels defined in DomainSum, we conduct a comprehensive analysis of the benchmark using eight key measures to examine the data distribution and summarization style shift characteristics at each level.

3.2.1 Summarization Domain Characteristic Measures

Given a set of document-summary paris (D_t, S_t) from a specific domain t, we employ the following detialed summarization domain characteristic measures. For n-grams, we select $n = 3$, which covers unigrams, bigrams, and trigrams in all measures.

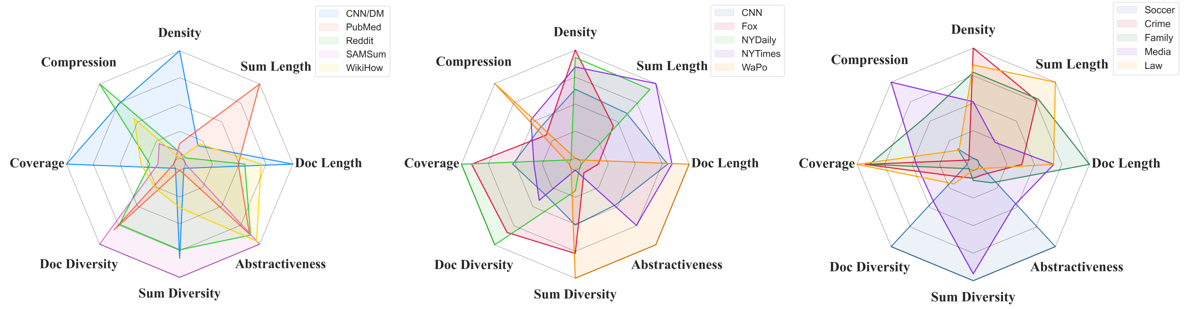
Length This measure includes both the average token length of documents $|D_t|$ and the average token length of summaries $|S_t|$.

Compression Compression measures the ratio of document length to summary length, indicating how much information has been condensed through summarization (Grusky et al., 2018).

$$\text{Compression}(D_t, S_t) = \frac{|D_t|}{|S_t|}. \quad (1)$$

Density Density measures the degree to which a summary is directly extracted from its corresponding document (Grusky et al., 2018).

$$\text{Density}(D_t, S_t) = \frac{1}{|S_t|} \sum_{f \in \mathcal{F}(D_t, S_t)} |f|^2, \quad (2)$$



(a) Genre Shift (b) Style Shift (c) Topic Shift

Figure 3: Radar charts illustrating multi-dimensional benchmark analysis based on domain characteristic measures across three levels. The measures evaluated include document/summary length, density, compression, coverage, abstractiveness, and document/summary diversity.

where $\mathcal{F}(D_t, S_t)$ is the set of extractive fragments shared between the document and the summary. It is formally defined as:

Diversity Diversity measures the variety of word usage within both the document and the summary (Yogatama et al., 2015). It is calculated separately for the document and the summary as the average proportion of unique n-grams. Document Diversity is calculated as:

$$\text{Document Diversity}(D_t) = \frac{1}{n}\sum_{i=1}^{n}\frac{|U_i(D_t)|}{|G_i(D_t)|}, \tag{3}$$

where $U_i(D_t)$ represents the set of unique n-grams in the document D_t , and $G_i(D_t)$ is the total set of n-grams in the document. Similarly, Summary Diversity is calculated as:

$$\text{Summary Diversity}(S_t) = \frac{1}{n}\sum_{i=1}^{n}\frac{|U_i(S_t)|}{|G_i(S_t)|}. \tag{4}$$

Coverage Coverage measures the proportion of n-grams from the document that appear in the summary (Yogatama et al., 2015).

$$\text{Coverage}(D_t, S_t) = \frac{1}{n}\sum_{i=1}^{n}\frac{|G_i(S_t) \cap G_i(D_t)|}{|G_i(S_t)|}. \tag{5}$$

Abstractiveness Abstractiveness measures the extent to which a summary diverges from direct extraction by employing novel words or phrases (Narayan et al., 2018). We calculate abstractiveness based on novel n-grams as:

$$\text{Abstractiveness}(D_t, S_t) = \frac{1}{n}\sum_{i=1}^{n}\frac{|N_i(S_t, D_t)|}{|G_i(S_t)|}, \tag{6}$$

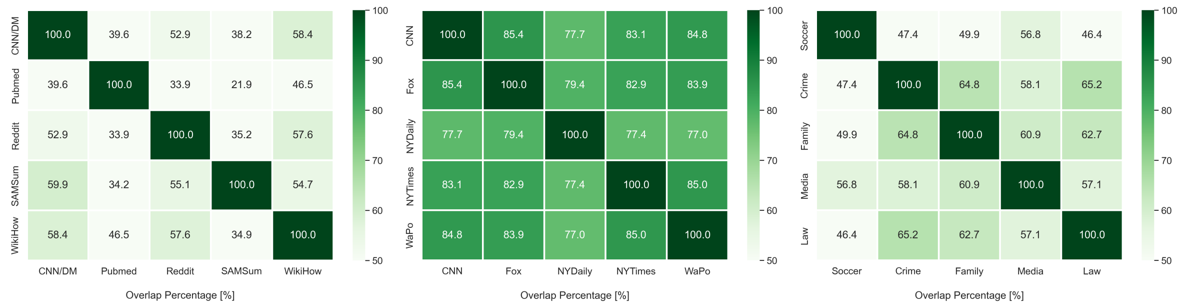
where $N_i(S_t, D_t)$ represents the set of novel n-grams in the summary that do not appear in the document D_t .

3.2.2 Data Distribution Analysis

Here we conduct analysis on DomainSum based on the above measures, and the overall results are presented in Figure 3.

Genre Level As shown in Fig 3a, CNN/DM exhibits the highest density and coverage, aligning with its news feature where summaries frequently retain substantial information from the original text. PubMed produces the longest summaries and achieves the highest abstractiveness, likely due to the complexity and specificity inherent in biomedical research articles. In contrast, Reddit demonstrates the highest compression ratio, reflecting the informal and concise nature of user-generated social post. SAMSum, which primarily comprises dialogues, demonstrates the greatest diversity in summaries, likely resulting from the varied and informal language structures typical of conversational data.

Style Level As illustrated in Fig 3b, news articles from different publishers exhibit markedly different styles. The Washington Post (WaPo) demonstrates the highest document length, compression, and abstractiveness, indicating that its summaries condense a substantial amount of content while introducing significant new material beyond the original text. Fox showcases the greatest summary diversity, reflecting a higher degree of linguistic variation in its summaries. In contrast, NYDaily ranks highest in coverage and density, as its summaries retain the most information directly from the original documents.



(a) Genre Shift

(b) Style Shift

(c) Topic Shift

Figure 4: Vocabulary overlaps of different levels of DomainSum . Vocabularies for each domain are created by considering the top 10K most frequent words (excluding stopwords).

Topic Level We further examine news articles from CNN covering different topics, as illustrated in Fig 3c. We observe that the Family category has the highest document length and compression, indicating that its summaries condense a substantial amount of content. The Crime category demonstrates the highest density, signifying that its summaries are closely aligned with the original text. The Law category shows the highest coverage, capturing more aspects of the original news in its summaries. Conversely, Soccer exhibits the highest abstractiveness, suggesting that its summaries introduce a greater amount of new information beyond the original text. Both the Media and Soccer categories display higher summary diversity, indicating a larger variety of language in their summaries.

Cross Level Comparison We also analyze the domain shift differences across levels. In terms of density, the Genre Shift shows high variability while the Style Shift displays more consistent density across categories, and the Topic Shift has a relatively uniform density distribution. Similarly for coverage, the Genre Shift demonstrates significant variation among categories, while the Style Shift presents a more evenly distributed coverage.

Overall, the Topic Shift shows the most consistent and balanced distribution across measures, while the Genre Shift displays the most extreme variations in density and coverage. The Style Shift falls in between, exhibiting moderate variations. These findings suggest that **topic shifts result in more uniform changes across text characteristics, whereas genre shifts lead to more dramatic variations, with style shifts producing intermediate effects**. This aligns with the **hierarchical design** of DomainSum and supports our hypothesis to categorize summarization domain shifts into hierarchical genre, style, and topic shifts.

3.3 Vocabulary Overlap

We also examine vocabulary overlaps between different domains across the three levels in DomainSum. As shown in Fig 4, the style shift shows the highest overlap between different news styles, consistently exceeding 75%, likely due to shared source characteristics. In contrast, the genre shift level has lower overlap percentages, mostly falling below 60%, indicating more distinct vocabulary usage across genres. The topic shift level reveals mixed overlap across news topics.

4 Experiment

4.1 Experimental Settings

Models We investigate popular PLMs and LLMs on DomainSum. For PLMs, we consider encoder-decoder models **BART** (Lewis, 2019) and **PEGASUS-X** (Phang et al., 2023). For LLMs, we include OpenAI's **GPT-4o mini** (OpenAI et al., 2024) and several top-ranked instruction-tuned open-source LLMs from Chatbot Arena (Zheng et al., 2023)[1], including META's **Llama3.1-8B-Instruct** and **Llama3.1-70B-Instruct** (Dubey et al., 2024), Mistral AI's **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023), and Google's **Gemma1.1-7B-Instruct** (Team et al., 2024). We use the official API for GPT-4o mini[2], and the corresponding Hugging Face APIs[3] for other LLMs.

We evaluated all LLMs in both zero-shot and few-shot ICL settings, and evaluate BART, PEGASUS-X, and Llama under fine-tuning settings with commonly used metrics including

[1] https://leaderboard.lmsys.org
[2] https://openai.com/index/openai-api/
[3] https://huggingface.co/docs/api-inference/index

	ROUGE	BERTScore	ROUGE	BERTScore	ROUGE	BERTScore	ROUGE	BERTScore	ROUGE	BERTScore
Genre Shift	**CNN/DM**		**PubMed**		**Reddit**		**SAMSum**		**WikiHow**	
GPT-4o mini	24.78 (-0.07)	87.43 (+0.06)	20.69 (+0.33)	83.60 (+0.14)	10.91 (+5.20)	81.48 (+1.58)	25.01 (+4.83)	89.34 (+1.08)	15.27 (+3.39)	83.20 (+1.27)
Mistral-7B	25.82 (-0.65)	87.43 (-0.10)	22.06 (-0.59)	83.49 (+0.28)	10.72 (+2.68)	81.36 (+0.91)	23.76 (+4.44)	89.10 (+0.94)	15.29 (+1.93)	83.09 (+0.80)
Gemma1.1-7B	23.68 (+0.66)	86.28 (+0.28)	20.74 (-1.21)	83.25 (+0.14)	11.50 (-0.02)	81.72 (+0.01)	23.26 (+2.92)	88.51 (+1.33)	16.09 (+0.66)	82.72 (+0.47)
Llama3.1-8B	26.02 (-0.80)	87.46 (+0.14)	21.07 (+1.54)	83.54 (+0.18)	11.83 (+3.82)	81.81 (+1.07)	26.11 (+5.05)	89.61 (+0.84)	16.75 (+1.69)	83.21 (+0.71)
Llama3.1-70B	25.87 (-0.13)	87.11 (-2.08)	21.71 (+1.85)	83.48 (-7.55)	11.99 (+4.90)	80.59 (+1.89)	25.48 (+4.69)	88.92 (+1.40)	17.05 (+2.93)	83.31 (+0.70)
Style Shift	**CNN**		**Fox**		**NYDaily**		**NYTimes**		**WaPo**	
GPT-4o mini	16.76 (+0.11)	84.58 (+0.11)	19.06 (-0.20)	85.27 (+0.17)	20.80 (+0.05)	85.19 (+0.34)	15.37 (-0.52)	84.32 (+0.16)	11.33 (+0.10)	83.52 (+0.18)
Mistral-7B	18.18 (-0.01)	84.64 (-0.13)	21.21 (-0.30)	85.42 (+0.04)	22.52 (-0.40)	85.33 (-0.02)	16.86 (+0.04)	84.45 (+0.05)	12.35 (+0.04)	83.52 (+0.11)
Gemma1.1-7B	15.94 (+0.52)	82.85 (+0.34)	18.92 (-0.20)	84.47 (+0.47)	20.19 (+0.35)	84.51 (-0.23)	14.93 (-0.46)	83.32 (+0.41)	10.89 (+0.22)	81.51 (+0.70)
Llama3.1-8B	17.46 (+1.68)	83.85 (-0.08)	19.95 (+2.75)	85.17 (+0.42)	21.30 (+2.83)	85.06 (+0.37)	16.19 (+0.41)	84.19 (-0.03)	11.75 (+0.71)	82.31 (+0.40)
Llama3.1-70B	17.06 (+3.63)	83.48 (-3.63)	19.24 (+6.96)	83.30 (+2.13)	20.24 (+3.49)	81.52 (-1.22)	15.77 (+2.45)	81.67 (+1.93)	11.60 (+1.85)	78.78 (+4.09)
Topic Shift	**Soccer**		**Crime**		**Family**		**Media**		**Law**	
GPT-4o mini	23.67 (+0.33)	86.96 (+0.20)	26.90 (+0.12)	87.92 (+0.09)	26.85 (-0.34)	87.78 (-0.01)	22.08 (-0.23)	86.95 (+0.02)	27.77 (+0.22)	87.95 (+0.03)
Mistral-7B	23.19 (+1.64)	81.58 (+5.45)	26.17 (+1.02)	83.86 (+3.85)	23.16 (+3.77)	87.04 (+0.52)	22.96 (-1.42)	85.29 (-4.05)	27.34 (+0.60)	86.25 (+1.45)
Gemma1.1-7B	23.65 (+0.08)	86.07 (+0.40)	25.89 (-0.11)	86.80 (+0.35)	25.06 (-0.74)	86.52 (+0.23)	21.10 (-0.37)	85.90 (+0.49)	25.53 (-0.05)	86.53 (+0.26)
Llama3.1-8B	25.06 (+1.07)	87.01 (+0.18)	27.95 (+1.22)	87.83 (+0.13)	27.80 (+0.29)	87.71 (-0.01)	23.28 (+0.60)	86.98 (+0.09)	28.48 (+0.75)	87.86 (+0.02)
Llama3.1-70B	24.53 (+1.12)	84.65 (-2.68)	27.33 (+1.48)	84.89 (+0.32)	28.31 (+0.45)	87.10 (+0.01)	21.73 (+2.06)	83.71 (+1.14)	28.59 (+1.01)	87.75 (-1.26)

Table 2: Results for zero-shot and two-shot prompting across genre, style, and topic shift levels in DomainSum. Values in parentheses indicate the performance difference between two-shot and zero-shot prompting. The best results for zero-shot prompting are highlighted in red, two-shot in blue, and cases where both achieve the best results are highlighted in purple.

BERTScore F1 scores (Zhang et al., 2020b) and the geometric mean of ROUGE-1/2/L scores (ROUGE) (Lin, 2004). Example zero-shot and few-shot ICL prompts can be found in Appendix B. More details on the fine-tuning settings can be found in Appendix C.

Data Sampling Due to the significant variation in the number of instances across datasets in DomainSum, we sample 10,000 instances for training, 500 for validation, and 500 for testing from the respective train, validation, and test sets of each dataset. All models are evaluated using these sampled test sets, with BART, PEGASUS-X and Llama3.1-8B fine-tuned on the corresponding sampled training sets. For LLMs, we also sample training sets of 20 and 1,000 instances for two-shot prompting and fine-tuning experiments. In the two-shot prompting setup, LLMs are prompted with two randomly selected instances from the training set. All sampled datasets and prompts used for ICL are released alongside our source code.

4.2 In-domain Evaluation

Zero/Few-Shot Results Table 2 presents the performance of models across genre, style, and topic levels in zero-shot and two-shot settings, evaluated using ROUGE and BERTScore metrics. At the genre level, Llama3.1-8B achieves the highest performance across most datasets, while Llama3.1-70B shows more substantial improvements when transitioning from zero-shot to two-shot. In the style shift setting, Mistral-7B performs best in zero-shot, but Llama3.1-70B outperforms it in two-shot. Finally, in the topic shift setting, Llama3.1-8B consistently outperforms other models in both zero-shot and two-shot settings, with GPT-4O mini achieving the best BERTScore on three datasets. Additionally, we observe that increasing the model size from Llama3.1-8B to Llama3.1-70B does not lead to significant overall performance improvements.

Fine-Tuning Results Table 3 presents the in-domain fine-tuning results. The results show that fine-tuned models generally outperform zero-shot and few-shot LLMs, underscoring the importance of in-domain training for this task. BART achieves the highest scores across all levels and datasets, except for slightly lower performance on PubMed. Llama3.1-8B also demonstrates significant performance improvements with fine-tuning. This is potentially due to its larger model size, which may require more data to fully benefit from in-domain training. The other possible explanation is that LLMs like Llama3.1-8B are pre-trained on more diverse and general-purpose data, making them less specialized for summarization. In contrast, smaller models like BART may benefit from more focused pretraining objectives and require fewer data to adapt effectively to a specific task. Moreover, fine-tuning a large model often demands a significantly larger dataset to fully leverage its capacity, whereas a smaller model can converge more efficiently given a limited amount of training data.

	ROUGE	BERTScore	ROUGE	BERTScore	ROUGE	BERTScore	ROUGE	BERTScore	ROUGE	BERTScore
Genre Shift	**CNN/DM**		**PubMed**		**Reddit**		**SAMSum**		**WikiHow**	
BART	28.43	87.72	20.70	84.05	20.86	87.46	37.61	91.56	24.73	87.62
PEGASUS-X	25.54	86.73	21.31	83.38	15.74	85.37	31.54	90.40	16.75	83.08
Llama3.1-8B	25.81	86.91	24.10	84.44	13.31	79.69	17.67	87.29	19.78	83.53
Style Shift	**CNN**		**Fox**		**NYDaily**		**NYTimes**		**WaPo**	
BART	42.02	89.44	55.05	91.55	53.84	91.06	25.38	86.81	20.05	85.58
PEGASUS-X	38.20	88.16	54.58	91.34	49.45	90.33	23.62	85.82	18.34	84.39
Llama3.1-8B	20.55	85.27	29.20	86.84	33.91	87.70	16.33	83.60	10.66	82.70
Topic Shift	**Soccer**		**Crime**		**Family**		**Media**		**Law**	
BART	29.71	88.90	31.45	88.98	30.45	88.67	27.56	88.20	31.50	88.86
PEGASUS-X	24.74	87.79	27.32	88.09	27.00	87.89	24.12	87.54	28.17	88.19
Llama3.1-8B	23.13	85.98	28.18	87.39	24.35	85.72	23.11	85.71	27.39	87.10

Table 3: Fine-tuning performance across five domains based on genre, style, and topic shift. Models are trained and evaluated within the same domain.

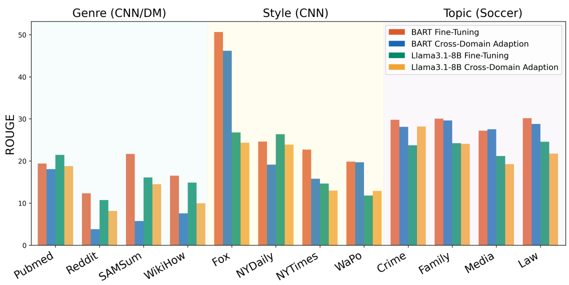


Figure 5: Comparison of in-domain fine-tuning and domain adaptation results for BART and LLaMA-3.1-8B models. The x-axis represents the target domains, with region colors indicating the source domains.

Cross-Level Results Comparison We also compare the performance differences of models across these three levels. As shown in Table 2 and Table 3, in the topic shift, models perform consistently well in both zero-shot and two-shot settings, suggesting that topic shifts are easier for models to generalize. In contrast, genre shift proves to be the most challenging, with models showing significant performance variations, particularly on datasets with specific summary characteristics, such as the long summaries in PubMed. Even with two-shot prompting, improvements remain limited, reflecting the more pronounced data distribution differences in genre shifts. Style shift falls in between, with models exhibiting moderate performance variations, corresponding to the intermediate complexity of its data distribution. These findings further support the hierarchical categorization of domain shifts in abstractive summarization.

4.3 Cross-domain Evaluation

Figure 5 compares the cross-domain adaptation performance, where BART and Llama3.1-8B are **trained on one domain and tested on other domains at the same level**. According to the results, the genre shift (from CNN/DM) poses the greatest challenge for both models, showing significant performance drops when tested on domains outside the training set. The style shift (from CNN) presents a more moderate challenge, with models showing smaller but still noticeable performance variability. Although style-based differences are easier to manage than genre shifts, models still struggle to adapt to varied writing styles. The topic shift (from Soccer) introduces the least variability in performance, suggesting that topic-based domain changes are easier for models to handle, as they tend to share more common contextual or thematic elements.

When comparing the models, the LLM (Llama3.1-8B) does not always outperform the PLM (BART) in absolute performance scores. In several cases, particularly after fine-tuning, BART achieves higher ROUGE and BERTScore values across different domains. This suggests that PLMs like BART can be more effectively fine-tuned for specific tasks, likely due to their smaller model size and more targeted pretraining, which enables them to adapt more precisely during fine-tuning. In contrast, Llama3.1-8B demonstrates better generalization potential, particularly in style and topic shifts, where its performance remains more stable across different domains. We infer that LLMs require more extensive or tailored fine-tuning to fully leverage their capacity, as their larger scale and broader pretraining may make fine-tuning less focused and efficient compared to PLMs.

Overall, the experimental results indicate that summarization domain shifts have hierarchical structures, with genre shifts being the most challenging for models to handle, followed by style shifts, while topic shifts lead to the most stable cross-domain performance.

Measure	Genre	Style	Topic
Doc Length	-0.305	-0.935*	0.275
Sum Length	-0.159	0.436	0.886*
Compression	-0.431	-0.915*	-0.680
Doc Diversity	0.022	0.944*	-0.617
Sum Diversity	-0.319	-0.998*	-0.929*
Coverage	0.313	0.994*	0.764
Abstractiveness	-0.321	-0.997*	-0.718
Density	0.424	0.830	0.155

Table 4: Pearson correlation coefficients between domain characteristic measures and model performance, averaged across model performance under three settings: zero-shot prompting, two-shot prompting, and fine-tuning. Asterisks (*) denote statistically significant correlations ($p \leq 0.05$).

4.4 Correlation Analysis

We present the Pearson correlation coefficients between our summarization measures and model performance across the genre, style, and topic dataset levels. As shown in Table 4, the correlation patterns differ significantly across these levels. At the genre level, none of the measures exhibit statistically significant correlations with model performance. In contrast, at the style level, several measures show significant positive or negative correlations ($p \leq 0.05$). Notably, summary diversity shows the strongest negative correlation and Coverage shows the strongest positive correlation with model performance. At the topic level, two measures exhibit significant correlations: summary length is positively correlated with performance, while summary diversity shows a significant negative correlation. The distinct correlation patterns across genre, style, and topic levels demonstrate that summarization performance is influenced by different aspects of the data distribution, and underscore the need for further investigation into how different summarization measures can be tailored to improve model performance across various domain shifts.

5 Conclusion

This paper presents DomainSum, a hierarchical benchmark for fine-grained domain shifts in abstractive summarization. We categorize domain shifts into genre, style, and topic levels and verify that genre shifts cause more variation, while topic shifts lead to more uniform changes through comprehensive multidimensional benchmark analysis. We also evaluate PLMs and LLMs on in-domain and cross-domain settings, showcasing their domain generalization abilities. DomainSum provides a valuable testbed for future research on summarization model generalization.

Acknowledgment

This work was enabled in part by funding from the National Science Foundation awards: 2201428, 2232862, 2004014, 2138259, and 2406251. We appreciate valuable suggestions from the anonymous reviewers.

Limitations

For the construction of our large-scale benchmark, we primarily reused existing datasets to represent various levels of domain shifts in DomainSum, rather than collecting new data, which would have been costly. The datasets we included are high-quality and publicly available for summarization tasks.

In our fine-tuning experiments, we sampled 10,000 training instances, 500 validation instances, and 500 test instances from the respective training, validation, and test sets of each dataset, rather than using the entire dataset. This decision was made to ensure fairness across datasets, as they vary in size. Previous research efforts (Goyal et al., 2022; Zhang et al., 2024b) have also tested GPT-3 on similarly small subsets.

Lastly, we did not include domain adaptation methods (Wang et al., 2019; Fabbri et al., 2020; Laskar et al., 2022) in our experiments. The main goal of this paper is to provide a benchmark for evaluating fine-grained domain shifts in summarization. Exploring techniques for adapting summarization models to new domains is left for future work, with our benchmark serving as a valuable resource for this purpose.

References

Anum Afzal, Ribin Chalumattu, Florian Matthes, and Laura Mascarell Espuny. 2024. Adapteval: Evaluating large language models on domain adaptation for text summarization. *arXiv preprint arXiv:2407.11591*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2017. Improving multi-document summarization via text classification. In *Proceedings of the AAAI conference on artificial intelligence*, 1.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

Chrysanne DiMarco and Graeme Hirst. 1993. A computational theory of goal-directed style in syntax. *Computational Linguistics*, 19(3):451–500.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Alexander R Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 2020. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. *arXiv preprint arXiv:2010.12836*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. 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, Volume 1 (Long Papers)*, pages 708–719.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Xinyu Hua and Lu Wang. 2017. A pilot study of domain adaptation effect for neural abstractive summarization. *arXiv preprint arXiv:1707.07062*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mahesh Joshi, Mark Dredze, William Cohen, and Carolyn Rose. 2012. Multi-domain learning: when do domains matter? In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1302–1312.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2018. Abstractive summarization of reddit posts with multi-level memory networks. *arXiv preprint arXiv:1811.00783*.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Xiangji Huang. 2022. Domain adaptation with pre-trained transformers for query-focused abstractive text summarization. *Computational Linguistics*, 48(2):279–320.

M Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinghao Li, Siyu Miao, Heyan Huang, and Yang Gao. 2024. Word matters: What influences domain adaptation in summarization? *arXiv preprint arXiv:2406.14828*.

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

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. Brio: Bringing order to abstractive summarization. *arXiv preprint arXiv:2203.16804*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ahmed Magooda and Diane Litman. 2020. Abstractive summarization for low resource data using domain transfer and data synthesis. In *The thirty-third international flairs conference*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

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

OpenAI, Josh Achiam, Steven Adler, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Jason Phang, Yao Zhao, and Peter J Liu. 2023. Investigating efficiently extending transformers for long input summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3946–3961.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp—a survey. *arXiv preprint arXiv:2006.00632*.

I Sutskever. 2014. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, et al. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Danqing Wang, Pengfei Liu, Ming Zhong, Jie Fu, Xipeng Qiu, and Xuanjing Huang. 2019. Exploring domain shift in extractive text summarization. *arXiv preprint arXiv:1908.11664*.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. *arXiv preprint arXiv:2305.13412*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Dani Yogatama, Fei Liu, and Noah A Smith. 2015. Extractive summarization by maximizing semantic volume. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1961–1966.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. Adaptsum: Towards low-resource domain adaptation for abstractive summarization. *arXiv preprint arXiv:2103.11332*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2022. Hegel: Hypergraph transformer for long document summarization. *arXiv preprint arXiv:2210.04126*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023a. Contrastive hierarchical discourse graph for scientific document summarization. In *The 4th workshop on Computa-tional Approaches to Discourse at the 61st Annual Meeting of the Association for Computational Linguistics (ACL CODI'23), July 9-14, 2023. Toronto, Canada.*

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023b. Diffusum: Generation enhanced extractive summarization with diffusion. *arXiv preprint arXiv:2305.01735*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023c. Extractive summarization via chatgpt for faithful summary generation. *arXiv preprint arXiv:2304.04193*.

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023d. Summit: Iterative text summarization via chatgpt. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10644–10657.

Haopeng Zhang, Philip S Yu, and Jiawei Zhang. 2024a. A systematic survey of text summarization: From statistical methods to large language models. *arXiv preprint arXiv:2406.11289*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024b. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*.

A Detailed Domain Characteristic Measure Results

The detailed measurement results in terms of all datasets included in DomainSum is presented in Table 5.

B Prompts

zero-shot prompt

> You are an expert at summarization. Summarize the following text: {document}
>
> Summary:

two-shot prompt

> You are an expert at summarization. Here are two examples of how to summarize a text:
>
> Example 1:
> Document: {doc_example1}
> Summary: {sum_example1}
>
> Example 2:
> Document: {doc_example2}
> Summary: {sum_example2}
>
> Now, summarize the following text:
>
> Document: {document}
>
> Summary:

The prompt used to fine-tune the Llama3.1-8B model.

> <|im_start|>user
> You are an expert at summarization. Summarize the following text: {document}
>
> Summary:<|im_end|>
> <|im_start|>assistant

C Detailed Fine-Tunning Settings

We fine-tune pre-trained BART and PEGASUS-X models using their base versions with huggingface packages (Wolf et al., 2020). We use AdamW optimizer (Loshchilov and Hutter, 2017) with learning rate set to 5×10^{-5}. We set the batch size to 12 and the number of training epoch to 4. The best model is selected based on the development set results. For Llama3.1-8B, we use the Low Rank Adaptation (LoRA) (Hu et al., 2021) fine-tuning method with huggingface packages, where the model is loaded to GPU as quantized 8-bit weights. We set the learning rate to 5×10^{-4}, the batch size to 4, and the number of training epoch to 20. All experiments are conducted on a single NVIDIA A100-SXM4-40GB GPU.

	Measures			Measures (%)				
Datasets	**Doc Length**	**Sum Length**	**Compression**	**Doc Diversity**	**Sum Diversity**	**Coverage**	**Abstractiveness**	**Density**
				Genre Shift				
CNN/DM	772.49	57.87	14.49	72.53	92.91	55.24	70.24	83.81
PubMed	444.00	209.51	2.32	78.86	80.50	27.55	87.80	19.44
Reddit	483.41	28.00	18.75	78.30	91.70	33.72	88.09	13.44
SamSum	126.59	23.09	5.82	80.34	95.58	31.62	90.52	13.85
WikiHow	582.18	62.19	11.27	74.58	85.76	35.98	89.68	9.31
				Style Shift				
CNN	971.18	30.92	36.91	75.00	96.40	63.42	52.14	10.77
Fox	654.45	28.69	33.47	77.80	96.77	75.34	35.11	14.50
NYDaily	577.10	34.97	20.65	78.71	95.96	78.32	32.78	13.80
NYTimes	990.96	35.97	41.38	75.43	95.69	57.22	63.16	12.89
WaPo	1069.26	22.89	60.09	73.08	97.09	46.94	73.50	4.21
				Topic Shift				
Soccer	642.51	52.81	12.61	75.96	93.31	48.35	78.45	30.65
Crime	730.59	66.30	12.11	71.31	91.97	58.07	67.37	94.20
Family	870.14	66.64	14.05	70.76	92.00	57.66	68.64	80.25
Media	796.58	56.81	15.60	73.07	93.22	53.19	72.12	63.25
Law	796.11	70.50	12.55	71.77	91.87	58.83	66.48	84.47

Table 5: Average Measures of domain adaptation summarization datasets. Length is represented in average tokens, Compression is a ratio, and diversity, coverage, abstractiveness, and density are expressed as percentages.

Model	Training Domain	Test Domain	ROUGE	BERTScore
BART	CNN/DM	Pubmed	19.39 (↓1.31)	84.03 (↓0.02)
		Reddit	12.33 (↓8.53)	84.23 (↓3.23)
		SAMSum	21.68 (↓15.93)	86.61 (↓4.95)
		WikiHow	16.39 (↓8.34)	85.07 (↓2.55)
	CNN	Fox	50.63 (↓4.42)	90.98 (↑0.57)
		NYDaily	47.53 (↓6.31)	90.18 (↓0.88)
		NYTimes	22.69 (↓2.69)	86.30 (↓0.51)
		WaPo	19.87 (↓0.18)	85.52 (↑0.06)
	Soccer	Crime	29.78 (↓1.67)	88.67 (↓0.31)
		Family	30.05 (↓0.40)	88.78 (↑0.11)
		Media	27.20 (↑0.36)	88.21 (↑0.01)
		Law	30.15 (↓1.35)	88.77 (↓0.09)
Llama3.1-8B	CNN/DM	Pubmed	21.45 (↓2.65)	82.43 (↓2.01)
		Reddit	10.73 (↓2.58)	80.50 (↑0.81)
		SAMSum	16.07 (↓1.60)	86.39 (↓0.90)
		WikiHow	14.87 (↓4.91)	82.56 (↓0.97)
	CNN	Fox	26.77 (↓2.43)	86.48 (↓0.36)
		NYDaily	26.36 (↓7.55)	86.26 (↓1.44)
		NYTimes	14.64 (↓1.69)	83.93 (↑0.33)
		WaPo	11.78 (↑1.12)	83.19 (↑0.49)
	Soccer	Crime	23.73 (↓4.45)	86.32 (↓1.07)
		Family	24.22 (↓0.13)	86.55 (↑0.83)
		Media	21.18 (↓1.93)	86.04 (↑0.34)
		Law	24.58 (↓2.81)	86.53 (↓0.57)

Table 6: Results of single-domain adaptation based on fine-tuned BART and Llama3.1-8B models. The numbers in parentheses represent the performance gain for cross-domain training compared to fine-tuning; a negative value indicates a performance decrease in cross-domain training.

D Experimental Results

The detailed ROUGE scores and BERTScores of BART and Llama3.1-8B for cross domain adaptation on three levels in DomainSum is shown in Table 6.