Word Matters: What Influences Domain Adaptation in Summarization?

Yinghao Li[1*] **Siyu Miao**[1*] **Heyan Huang**[12†] **Yang Gao**[12†]
[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[2]Beijing Institute of Technology Southeast Academy of Information Technology, Putian, China
{yhli,symiao,hhy63,gyang}@bit.edu.cn

Abstract

Domain adaptation aims to enable Large Language Models (LLMs) to generalize domain datasets unseen effectively during the training phase. However, factors such as the size of the model parameters and the scale of training data are general influencers and do not reflect the nuances of domain adaptation performance. This paper investigates the fine-grained factors affecting domain adaptation performance, analyzing the specific impact of 'words' in training data on summarization tasks. We propose quantifying dataset learning difficulty as the learning difficulty of generative summarization, which is determined by two indicators: word-based compression rate and abstraction level. Our experiments conclude that, when considering dataset learning difficulty, the cross-domain overlap and the performance gain in summarization tasks exhibit an approximate linear relationship, which is not directly related to the number of words. Based on this finding, predicting a model's performance on unknown domain datasets is possible without undergoing training. Source code and scripts are available at https://github.com/li-aolong/Word-Matters.

1 Introduction

With the continuous development of Large Language Models (LLMs), remarkable capabilities have been demonstrated in knowledge comprehension (Thirunavukarasu et al., 2023a; Sun et al., 2023), logical reasoning (Hao et al., 2023; Miao et al., 2023), problem-solving (Chan et al., 2023; Talebirad and Nadiri, 2023), and other aspects (Zhao et al., 2023; Wen et al., 2023).

As a result, LLMs have been widely applied to various summarization tasks on different domains to improve productivity like law (Shukla et al., 2022), medicine (Veen et al., 2023), finance (Li et al., 2023) and so on, including both natural and social science study (Glickman and Zhang, 2024; Xu et al., 2024). However, when LLMs are applied to specific domains, it often necessitates the selection of corresponding domain-specific knowledge bases for training (Zhang et al., 2023; Cui et al., 2024; Thirunavukarasu et al., 2023b; Yu et al., 2021). This results in a limitation where a model trained in one domain struggles to be effectively applied in others (Dada et al., 2023), leading to a waste of resources.

This constraint arises from the disparity in the distribution between the training data and the target domain data (Zhang et al., 2022). In light of this limitation, effective methods must be taken to fix the gap and then enhance the model's adaptability and efficiency in summarization tasks. Domain adaptation aims to train a model from multiple source domains, enabling it to generalize well to unseen domains (Li et al., 2018; Dou et al., 2019). Consequently, enhancing domain adaptation performance is a key objective for large-scale models in improving downstream tasks (Zhou et al., 2021a). It is worthwhile to explore which factors can affect the domain adaptation performance (Wang et al., 2021).

Schaeffer et al. (2023) proposes that metrics based on nonlinear or non-contiguous tokens are crucial to a model demonstrating emergent abilities and that ROUGE-L-Sum shows sharper variations. This has inspired us to consider the performance changes of models in domain adaptation from the perspective of more granular units. Tokens typically do not possess complete semantics, whereas words are the basic language units with specific meanings or functions. Therefore, we consider exploring the impact on model performance in domain adaptation from the perspective of words.

Summarization tasks involve generating concise texts that encapsulate the main components of longer documents, considering factors such

[*]These authors are both First Author.
[†]Corresponding author

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13236–13249

as coherence, information diversity, and coverage scope (Alomari et al., 2022). This differs from other downstream tasks like machine translation (Klimova et al., 2022) and classification (Bird et al., 2020). Fatima et al. (2022) note that reducing summary extractors' size or compression ratio can lead to losing vital content, features, concepts, and other significant information. Therefore, we explore how the degree of information extraction between input documents and target summaries impacts domain adaptation performance in summarization tasks.

This paper investigates how words impact the domain adaptation of summary tasks. We first introduce two indicators, compression rate, and abstraction level, to quantify the learning difficulty of datasets, thereby more accurately reflecting the performance gain of models. Then, we identify two key aspects affecting domain adaptation: cross-domain overlap and word count, hypothesizing a linear relationship between them and model performance. Experiments are conducted with models of various sizes on summarization datasets from four domains. The results indicate that the cross-domain overlap exhibits an approximately linear relationship with performance gain when considering dataset learning difficulty. In contrast, word count shows no significant correlation. Based on this linear relationship, it is possible to predict model performance without undergoing training by using the cross-domain overlap calculated from the dataset. Our contributions can be summarized as follows:

- We propose two factors affecting the domain adaptation of summarization tasks: (1) **Learning difficulty coefficient** of the dataset more accurately reflects the performance gain; (2) **Cross-domain overlap** directly represents the closeness between the source and target domains.
- Our experiments show that cross-domain overlap has an approximately linear relationship with performance gains based on the learning difficulty coefficient, revealing the connection between datasets and domain adaptation from the perspective of words.
- We demonstrate that without undergoing training, it is possible to predict a model's performance on unknown domain datasets solely based on the learning difficulty coefficient and cross-domain overlap. This provides a resource-efficient and rapid validation method for models regarding domain adaptation.

2 Related Work

Domain Adaptation (DA) has emerged as a crucial methodology for enhancing model performance across varying domains (Farahani et al., 2020). DA aims to enhance the performance of LLMs in a target domain, where annotated data may be scarce or absent, by leveraging knowledge from a related domain with a sufficient amount of labeled data (Farahani et al., 2020). Many methods have been developed to tackle the out-of-domain adaptation issue (Zhou et al., 2021b; Fan et al., 2021; Cha et al., 2021; Wang et al., 2022; Ling et al., 2023). There are three pivotal strategies related to our work: (1) Continual pre-training, (2) Alignment of distributions, and (3) Adaptation tuning.

Continual Pre-training Continual pre-training uses similar training objectives as continual self-supervised training to update pre-trained models with new data instead of retraining from scratch (Gupta et al., 2023). Continual pre-training is studied for domain adaptation where the new dataset comes from a new domain, which is referred to as continual domain-adaptive pre-training (DA-training) (Gururangan et al., 2021; Scialom et al., 2022; Ke et al., 2023a). DAP-training methods can achieve better results by training LLMs with a large unlabeled domain corpus before end-task fine-tuning (Alsentzer et al., 2019; Lee et al., 2019; Gururangan et al., 2020; Ke et al., 2023b). However, the effectiveness of this method is contingent upon the relevance of the pre-trained LLMs to the target domain and requires substantial domain-specific data to achieve optimal performance.

Alignment of Distributions Aligning the statistical attribution of the source and target domains to match their distributions has emerged as a principal method (Peng et al., 2019; Nguyen-Meidine et al., 2020). The general way of aligning the distributions is by minimizing the distance between domains. The most used distance measures in domain adaptation are maximum mean discrepancy, Kullback-Leibler divergence, and contrastive domain discrepancy (Long et al., 2015; Ganin and Lempitsky, 2015). The strength of this approach lies in its theoretical rigor and the potential for precise domain alignment. Nevertheless, the challenge

of selecting appropriate distance metrics and the computational complexity of these calculations can pose significant obstacles. Compared with this, we adopt word-based statistical metrics to calculate the similarity between texts from different domains directly.

Adaptation Tuning LLMs may not capture sufficient knowledge for specific tasks or domains even when trained on vast amounts of general text data. Adapting models to a smaller, domain-specific dataset can significantly improve their performance in that specific area. Here are three primary methods for adapting LLMs: (1) Prompt engineering has shown its power to quickly adapt LLMs to unseen domains without updating the inner parameters. Prompts define unseen tasks with or without several illustrative examples to LLMs in natural language (Ben-David et al., 2022; Kojima et al., 2023). Continuous prompts are sequences of tokens attached with the input sentence that can be learned from the downstream dataset by prompt tuning (Ye et al., 2022; Vu et al., 2022; Razdaibiedina et al., 2023). Su et al. (2022) demonstrate the transferability of continuous prompts in both cross-task as well as cross-model settings; (2) Adapter fine-tuning, such as Low-rank adapters (Hu et al., 2021) and DyLora (Valipour et al., 2023), adds a small number of extra parameters to LLMs to enhance performance without major modifications; (3) Full fine-tuning is still the most fundamental and wildly used method to improve the model's adaptation performance. Instruction fine-tuning has proven to be highly successful in enhancing the model's adaptation capabilities (Chung et al., 2022; Menick et al., 2022; Wei et al., 2022; Huang et al., 2023). However, how to select suitable data to cultivate LLMs' adapting capacity and predict transferring results remain a problem.

3 What and How does Word Influence Domain Adaptation?

In this section, we explore how words can affect aspects of domain adaptation and their impact. We first hypothesize that datasets of varying learning difficulties affect model performance and investigate the influence of words on the learning difficulty of target domain datasets. We propose two indicators to reflect the learning difficulty of datasets: Compression Ratio and Abstraction Level. Secondly, from the perspective of words, we propose two aspects that could affect domain adaptation: cross-domain overlap and word count. Finally, we hypothesize a linear relationship between these aspects and domain adaptation performance based on dataset learning difficulty. This hypothesis is tested in subsequent experiments.

3.1 Word Influence On Target Domain Dataset Learning Difficulty

We assume that different datasets have varying levels of learning difficulty, and training models on datasets with low learning difficulty can lead to higher metric improvements on the test set. In contrast, the metric improvement is relatively small for datasets with high learning difficulty. To quantitatively assess the learning difficulty of datasets for the generative summarization task, we introduce two indicators: Compression Ratio and Abstraction Level.

Compression Ratio The Compression Ratio reflects the learning difficulty of a dataset in terms of form, which describes the degree of length reduction of the original text relative to the generated text. A higher compression Ratio indicates a more challenging dataset because the model needs to compress the content of the original documents to a greater extent, which places a higher demand on the model's text comprehension and information extraction capabilities. The Compression Ratio α for a dataset containing n samples is represented as the average Compression Ratio across all samples and is calculated using the following formula:

$$\alpha = \frac{1}{n}\sum_{i=1}^{n}\frac{|D_i|}{|S_i|}, \tag{1}$$

where $|D_i|$ and $|S_i|$ represent the word count of the i-th document and summary in the dataset, respectively.

Abstraction Level Considering only the Compression Ratio cannot fully reflect the dataset's learning difficulty. For example, if the reference summaries in the dataset are verbatim excerpts of specific sentences from the source document, even though the Compression Ratio may be high, the model only needs to learn to copy parts of the document to achieve high performance. This is a straightforward extractive pattern and does not truly reflect the model's summarization capability. Therefore, we introduce Abstraction Level that reflects the learning difficulty of a dataset in terms of content, which we define as the reciprocal of

the average ROUGE score between the original documents in the test set and the corresponding summaries, with the formula represented as follows:

$$\beta = \frac{n}{\sum_{i=1}^{n} ROUGE_{(d_i, s_i)}} \quad (2)$$

ROUGE is essentially a method for calculating overlap. We argue that the overlap between documents and summaries can, to some extent, represent the co-occurrence of knowledge within the dataset. A lower ROUGE value indicates lower content relevance between the document and the summary, making improving performance on that dataset. Hence, we use the reciprocal of Abstraction Level to reflect the learning difficulty of the dataset in terms of content.

Learning Difficulty Coefficient According to the proposed two indicators influencing dataset learning difficulty, we define a dataset's learning difficulty coefficient λ as the product of Compression Ratio and Abstraction Level, represented by the following formula:

$$\lambda = \alpha\beta \quad (3)$$

3.2 Possible Impact Aspects Cross Different Domains Based on Words

We investigate the influence of words on domain adaptation performance. Due to the limitations of the original metric for summarization tasks, such as ROUGE, in reflecting how well a model generalizes across different domains, we introduce an evaluation metric suitable for assessing domain adaptation performance and explore the potential factors that might affect this metric.

Summarization Gain The commonly used evaluation metric in summarization tasks, ROUGE, reflects performance on a specific dataset, whereas the performance of domain adaptation is more evident in the change in absolute performance. Therefore, we employ the ROUGE gain as a fundamental measure of domain adaptation performance, with the formula as follows:

$$Gain = ROUGE_{\textit{fine-tuned}} - ROUGE_{base}, \quad (4)$$

where $ROUGE_{base}$ represents the original model's ROUGE value calculated through direct inference on the test set, while $ROUGE_{\textit{fine-tuned}}$ represents the ROUGE value obtained after the model has been fine-tuned.

Cross-domain Overlap When considering performance adaptation across different domains, intuitively, they are more similar if there are more overlapping words between datasets from different domains. We assume this similarity can lead to performance improvement in domain adaptation. We propose cross-domain overlap to characterize the word-level overlap ratio between different domains. For source domain S containing n datasets and target domain T containing m datasets, the formula for cross-domain overlap is expressed as follows:

$$\gamma = \frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\sum_{k=1}^{l} Count(w_k^{T_j}, S_i)}{|S_i|} \quad (5)$$

where S_i and T_i represent the i-th dataset for the domain S and T respectively, l is the total number of unique words in T_j, $w_k^{T_j}$ represents the k-th unique word in T_j, and $Count(w_k^{T_j}, S_i)$ is the number of occurrences of word $w_i^{T_j}$ in S_i.

Word Count Word count refers to the total number of words across all samples in a dataset. Generally, the larger the number of training set samples within a specific range, the better the model performance. However, whether a higher total word count in all samples is also beneficial is worth exploring. Therefore, we consider word count as one aspect affecting the model's domain adaptation performance.

3.3 How to influence?

We posit that the metrics of cross-domain overlap and Word Count notably influence domain adaptation performance, particularly when accounting for dataset learning difficulty. The higher the cross-domain overlap, the more similarity between the source domain's data and the target domain's data. Consequently, the model is more likely to leverage data from source domains to learn knowledge that is closer to the target domain, resulting in better adaptation to that domain. Regarding Word Count, it is intuitive to assume that a larger training dataset size, and consequently a higher Word Count, leads to better model performance. However, whether this correlation consistently extends to domain adaptation performance remains an open question for investigation.

Dataset	Domain	Training Set	Test Set	Document Words (Avg.)	Summary Words (Avg.)
CNNDM	News	35,000 (287113)	500 (11490)	663 (781)	56 (56)
PubMed	Science	35,000 (119,924)	500 (6,658)	1064 (3,049)	190 (202)
SAMSum	Conversation	14,732 (14,732)	500 (819)	93 (92)	20 (20)
WikiHow	General	35,000 (157,252)	500 (5,577)	523 (580)	52 (62)

Table 1: The statistics of datasets. The data in parentheses represent the total data from the original datasets.

LD-Gain When dataset learning difficulty is factored into domain adaptation performance, for more challenging datasets, a model should require a smaller performance gain to achieve the same level of performance as when difficulty is not considered since the dataset's complexity impedes the model's ability to generalize across domains. Hence, we hypothesize that cross-domain overlap and Word Count are linearly related to the product of dataset learning difficulty and performance gain, which is referred to **LD-Gain**. The proposed hypotheses are as follows:

$$\begin{aligned} &Hypothesis1 : \gamma \propto \lambda Gain = LD\text{-}Gain, \\ &Hypothesis2 : WC \propto \lambda Gain = LD\text{-}Gain, \end{aligned} \tag{6}$$

where WC represents Word Count. The following experiments section will verify the two hypotheses.

4 Experiments

We use four summarization datasets, CNN/Daily Mail (Hermann et al., 2015), PubMed (Cohan et al., 2018), SamSum (Gliwa et al., 2019) and WikiHow (Koupaee and Wang, 2018), each originating from the news, science, conversation, and general domains respectively. Due to the significant differences in the number of samples across different datasets, we sample 35,000 samples from the CNNDM, PubMed, and WikiHow datasets for the training set while retaining the entire SAMSum dataset. We sample 500 test samples from all the test sets of these datasets. The detailed statistical data of the datasets are shown in Table 1.

To verify the above two hypotheses, we configure different experimental setups for cross-domain overlap and Word Count. All experiments are based on the Bloom (Scao et al., 2022) and Llama2 (Touvron et al., 2023) series of models. Due to resource constraints, different experiments use models of various sizes. The prompts used to train the Llama2 and Bloom models are presented in appendix A.

4.1 Setup for Cross-domain Overlap

Cross-domain overlap is calculated between the different source and target domains. Considering the potential impact of the number of source domains on the results, we set up two experiments: single-domain adaptation and multi-domain adaptation.

Single-domain Adaptation Single-domain adaptation refers to training on a single source domain and testing on different target domains. We first test the basic performance of models on test sets across four domains and then fine-tune the models using training sets from the three domains, excluding the test set domain. For a model, this results in having models trained on three single-domain datasets, which are then tested for their performance on the test sets. Finally, we calculate the change in performance. The Bloom-1.1B, Bloom-3B, and Llama-2 7B models are trained on 4 RTX 3090 GPUs in this experiment. For the Bloom-1.1B and 3B models, we conduct full-parameter fine-tuning for one epoch with a learning rate 2e-5 and a batch size of 4. For the Llama2-7B model, we use LoRA (Hu et al., 2021) for fine-tuning over three epochs, with the other hyperparameters remaining unchanged.

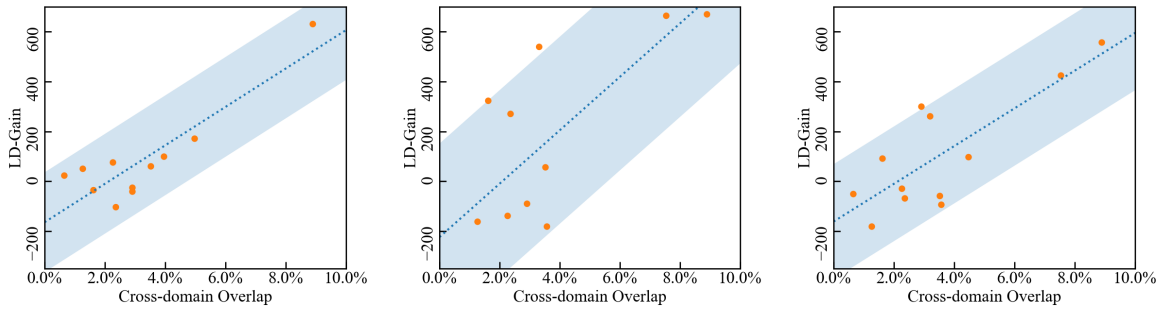
Multi-Domain Adaptation Multi-domain adaptation refers to training on multiple source domains. When one of the four domain datasets is selected as the test set, the datasets from the other three domains are combined to serve as the training set. In this experiment, we use the Bloom-1B and 3B models. The basic performance of the two models on the four domain test sets has already been tested in the single-domain adaptation, so we use the mixed dataset from the three domains, excluding the test set domain, for training to calculate the performance change. The training hyperparameters are the same as those used in the single-domain adaptation.

4.2 Setup for Word Count

To minimize the interference of other factors in investigating the relationship between word count and domain adaptation, we use only one domain dataset, CNNDM, for training. Then, we test on the same domain's test set and the test sets of the other three domains to observe the impact of word count on the results. The CNNDM training set is evenly divided into ten chunks, each used for training in separate stages. Training is conducted in ten stages, starting with the first chunk as the training set. In

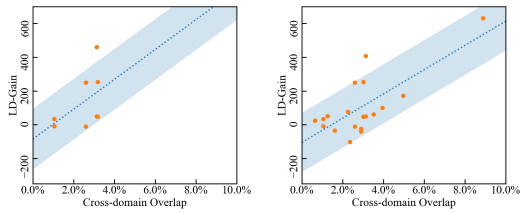
Training Set	Test Set	Compression Rate	Abstract Level	Learning Difficulty Coefficient	ROUGE Base	ROUGE Fine-tuned	ROUGE Improvement	Cross-domain Overlap	*LD-Gain*
PubMed						7.84	-0.26	2.35%	-68.08
SAMSum	CNNDM	12.95	20.22	261.85	8.10	10.23	2.13	8.89%	557.74
WikiHow						8.47	0.37	4.47%	96.88
CNNDM						10.65	3.28	2.90%	304.91
SAMSum	PubMed	7.08	13.13	92.96	7.42	6.80	-0.62	3.51%	-57.64
WikiHow						6.42	-1.00	3.56%	-92.96
CNNDM						6.36	1.14	1.62%	92.85
PubMed	SAMSum	4.86	16.76	81.45	5.22	4.61	-0.61	0.64%	-49.68
WikiHow						2.99	-2.23	1.26%	-181.63
CNNDM						5.075	0.513	3.30%	262.63
PubMed	WikiHow	13.05	39.23	511.95	4.562	4.506	-0.056	2.25%	-28.67
SAMSum						5.39	0.83	7.53%	424.92

Table 2: Results of single-domain adaptation on the Llama-2 7B model.



(a) Bloom-1.1B (b) Bloom-3B (c) Llama2-7B

Figure 1: Single-domain adaptation. The dotted line reflects a changing trend fitting the scatter data. The light-colored area represents a standard deviation of plus or minus.



(a) Multi-Domain (b) Mixed-Domain

Figure 2: The left is the result of multi-domain adaptation for Bloom-1.1B and 3B. The right contains the results of both single-domain and multi-domain adaptation.

subsequent stages, each phase uses the training set from the previous stage plus the next chunk, continuing until the final stage, where the entire dataset of the source domain is used for training. Each stage involves training for one epoch.

5 Results and Analysis

We analyze the results from different perspectives. Both single-domain and multi-domain experiment results supported our hypothesis that cross-domain overlap is linearly correlated with performance. Due to this finding, we could make a quantitative prediction about the transferred result of certain given training and evaluating datasets. On the other hand, we found that word count is unrelated to domain adaptation, which further emphasizes the significance of choosing high-quality data rather than a large amount of data.

To solidify the universality of our findings, we conduct a prediction study and it turns out that our hypothesis can successfully predict domain adaptation performance with cross-domain overlap. There is only a slight gap between our prediction data and observed data.

5.1 Cross-domain overlap is linearly correlated with performance

Single-Domain Adaptation The results of the single-domain adaptation for Llama2-7B are shown in Table 2, and the results of Bloom-1.1B and 3B are shown in Table 5 and Table 6 of Appendix B. The compression rate, abstraction level, and the learning difficulty coefficient calculated based on them only relate to the test set. Therefore, these indicators remain constant for the same test set,

Model	CNNDM	PubMed	SAMSum	WikiHow	Compression Rate	Abstract Level	Learning Difficulty Coefficient	ROUGE base	ROUGE Fine-tuned	ROUGE Improvement	Cross-domain Overlap	*LD-Gain*
Bloom-1B	●	○	○	○	12.95	20.22	261.85	3.72	3.91	0.19	3.13%	49.75
	○	●	○	○	7.08	13.13	92.96	4.13	4.65	0.52	3.18%	48.34
	○	○	●	○	4.86	16.76	81.45	1.75	1.65	-0.1	1.05%	-8.15
	○	○	○	●	13.05	39.23	511.95	2.51	2.49	-0.02	2.60%	-10.24
Bloom-3B	●	○	○	○	12.95	20.22	261.85	4.06	5.48	1.42	3.13%	371.83
	○	●	○	○	7.08	13.13	92.96	4.84	6.87	2.03	3.18%	188.71
	○	○	●	○	4.86	16.76	81.45	1.81	2.17	0.36	1.05%	29.32
	○	○	○	●	13.05	39.23	511.95	2.59	3	0.41	2.60%	209.90

Table 3: Multi-domain Adaptation. ● represents the test sets, while ○ constitutes the training sets together.

even if the training set changes. Similarly, the base ROUGE value of the model on the test set also remains unchanged.

Figure 1 illustrates the relationship between cross-domain overlap and LD-Gain for three different models. It can be observed that when there is a low overlap between the target domain and the source domain, the model fails to generalize well to the target domain. Conversely, the model performance tends to exhibit significant improvement when there is high vocabulary overlap. We also utilize another similarity calculation method, BERTScore (Zhang et al., 2019), to replace ROUGE values in computing LD-Gain. The results for single-domain adaptation on bloom-3b and llama2-7b are depicted in the figure 5 of Appendix C. We observe that the relationship between LD-Gain computed based on BERTScore and cross-domain overlap is similar to the results obtained with ROUGE-based calculations. Based on this discovery, we believe that there exists a linear correlation between vocabulary overlap and model performance gain, factoring in dataset learning difficulty.

Multi-Domain Adaptation Table 3 presents the results obtained by training with multiple domain data and testing with a single domain. It can be observed that the ROUGE improvement of Bloom-1B on CNNDM is 0.19, which is smaller than 0.52 compared to the improvement on PubMed. However, the learning difficulty coefficient for the CNNDM test set is higher at 261.85, exceeding the value of 92.96 for PubMed. Therefore, the LD-Gain of the model on CNNDM, adjusted by the learning difficulty coefficient, is higher than that of PubMed.

The relationship between cross-domain overlap and LD-Gain for multi-domain adaptation is illustrated in Figure 2a. We also combine the results of single-domain and multi-domain adaptation and plot them in a single graph, as shown in Figure 2b. It can be observed that as the cross-domain overlap increases, the model's actual gains gradually increase. This finding reveals that cross-domain overlap has a linear relationship with performance in domain adaptation.

5.2 Word count is not related to performance

The word count results are presented in Table 4. It can be observed that the word count within a chunk is relatively similar, indicating that the word count can be controlled by increasing the number of chunks. On the other hand, the overlap within a domain does not vary significantly. It remains stable within a small range, allowing for the observation of the relationship between overlap and performance across different domains.

The visual results from Figure 3 demonstrate that there is no clear upward or downward trend in model performance across different domains as the word count increases. Instead, there is oscillation within a specific range. The fluctuations are most prominent for the target domains CNNDM and SAMSum, which remain relatively stable within a specific range. Consequently, we conclude that the influence of word count on domain adaptation performance is unrelated. There are instances where performance may even decline with an increase in word count.

Meanwhile, we also calculate the overlap of each chunk to observe whether there is still a linear relationship between the improvement of model performance and overlapping data at different domains. Although the cross-domain overlap is relatively similar within the same domain, a linear correlation exists between cross-domain overlap and LD-Gain across different domains. Specifically, the model exhibits higher actual gains as the cross-domain overlap increases.

5.3 Predictability

The preceding experiments have confirmed a linear correlation between the cross-domain overlap

	Test Set	Chunk0	Chunk1	Chunk2	Chunk3	Chunk4	Chunk5	Chunk6	Chunk7	Chunk8	Chunk9
Word Count		2,608,429	2,594,348	2,591,597	2,573,224	2,641,309	2,558,078	2,582,793	2,594,890	2,607,471	2,614,820
Cross-domain Overlap	CNNDM	7.37%	7.35%	7.39%	7.35%	7.28%	7.29%	7.33%	7.28%	7.25%	7.34%
	PubMed	2.93%	3.00%	2.94%	2.87%	2.91%	2.92%	2.92%	2.95%	2.93%	2.84%
	SAMSum	1.59%	1.56%	1.74%	1.57%	1.60%	1.55%	1.60%	1.60%	1.64%	1.76%
	WikiHow	3.58%	3.57%	3.64%	3.60%	3.56%	3.55%	3.54%	3.55%	3.48%	3.55%
LD-Gain	CNNDM	851.00	775.07	971.45	735.79	960.98	995.02	1091.91	811.73	976.69	1099.76
	PubMed	43.69	31.61	21.38	26.96	33.47	49.27	25.10	26.03	35.32	40.90
	SAMSum	6.52	0.81	13.03	-24.44	-1.63	-24.44	-20.36	-8.96	-38.28	-17.11
	WikiHow	-102.39	-15.36	25.60	-66.55	-30.72	30.72	112.63	-35.84	20.48	-66.55

Table 4: Results of different chunks on test sets across various domains. The word count is increasing from Chunk0 to Chunk9.

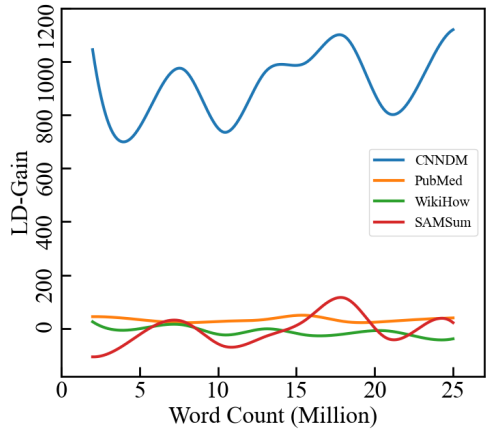


Figure 3: Relationship between word count and LD-Gain. As word count increases, there is no significant trend in LD-Gain change.

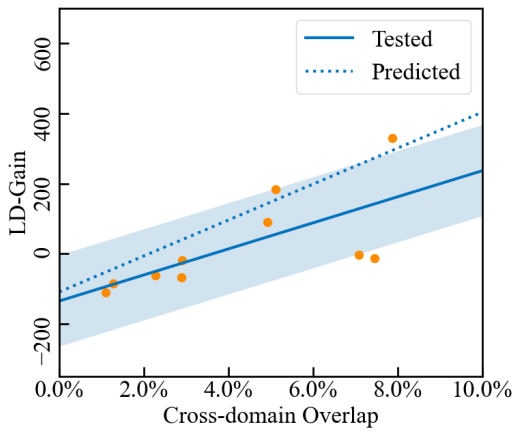


Figure 4: The dotted line is the predicted line from the previous experiment, while the solid line is drawn with the testifying dataset.

and performance, taking into account dataset learning difficulty. Based on this observation, we can extrapolate performance predictions for unknown domain datasets using the results from existing domain datasets. As depicted in Figure 1a, a linear fit to the scatter plot of single-domain data yields a performance prediction trend for Bloom-1.1B in single-domain adaptation, as illustrated by the dashed line in Figure 4.

We re-sample 500 different examples, distinct from the previous datasets. Subsequently, we compute the compression ratio and abstraction level values for the four new test sets, obtaining the learning difficulty coefficient λ. Based on the metrics from the training set used in the previous single-domain experiments, we calculate the cross-domain overlap γ value. Ideally, we can use λ and γ to predict the model's performance gain on the new dataset, thereby obtaining the predicted ROUGE value with the formula as follows:

$$\begin{aligned} ROUGE_{predicted} &= Gain + ROUGE_{base}, \\ Gain &= \frac{LD\text{-}Gain}{\lambda}, \\ LD\text{-}Gain &= \beta_0 + \beta_1 \gamma, \end{aligned} \tag{7}$$

where β represents the parameters of the fitting line under the existing data of the model.

We conduct experiments using the Bloom-1.1B model. The results of the new dataset, along with the fitted curve, are shown in Figure 4 as scatter points and a solid line. It can be observed that the predicted fitted line and the actual experimental results' fitted line exhibit a similar trend, with a relatively small difference. This indicates that we can estimate the performance for unseen domain datasets based on the dataset's characteristics and existing performance results, thereby obtaining a rough performance expectation before actual inference.

6 Conclusion

We investigate the impact of words on domain adaptation performance in summarization tasks. We propose two indicators to represent the learning difficulty from a dataset and introduce a performance evaluation method based on learning difficulty. We find that word overlap is an essential factor affecting domain adaptation and exhibits a linear correlation with model performance. However, the influence of word count on domain adaptation does not show a regular pattern. We will investigate this phenomenon further in future work.

Furthermore, by predicting the performance of new domain data based on its cross-domain overlap with existing domains, it becomes possible to preemptively assess the model's suitability for specific domains without the need for extensive retraining or fine-tuning. This predictive capability can significantly streamline the adaptation of language models to new domains and save many resources, finally improving their practical utility in real-world applications.

7 Limitations

The approaches of domain adaption mainly involve continual pre-training, alignment of distributions, and adaptation tuning. Our findings are related to the third one and, therefore, limited to discussing the influence factors of pre-training data, model parameters, and so on. For the adaptation tuning method, our paper focuses on exploring word impact on domain adaptation. Other factors, such as the quality of summarization training data, instruction diversity, and quality, are out of our consideration and may bring additional noise.

Due to resource constraints, this paper employed LoRA fine-tuning for a 7B model without investigating the effects of different fine-tuning methods on domain adaptation. In future work, we will explore in more detail the impact of dataset quality and different training methods on domain adaptation.

Acknowledgements

We appreciate Xiaochen Liu for providing the initial inspiration for this work. This work was supported by the Joint Funds of National Natural Science Foundation of China (No. U21B2009), Major Research Plan of the National Natural Science Foundation of China (Grant No. 92370110).

References

Ayham Alomari, Norisma Idris, Aznul Qalid Md Sabri, and Izzat Alsmadi. 2022. Deep reinforcement and transfer learning for abstractive text summarization: A review. *Computer Speech & Language*, 71:101276.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical bert embeddings.

Eyal Ben-David, Nadav Oved, and Roi Reichart. 2022. Pada: Example-based prompt learning for on-the-fly adaptation to unseen domains.

Jordan J. Bird, Anik'o Ek'art, and Diego Resende Faria. 2020. Chatbot interaction with artificial intelligence: human data augmentation with t5 and language transformer ensemble for text classification. *Journal of Ambient Intelligence and Humanized Computing*, 14:3129–3144.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. 2021. Swad: Domain generalization by seeking flat minima. In *Neural Information Processing Systems*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shan Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *ArXiv*, abs/2308.07201.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

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. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Jiaxi Cui, Munan Ning, Zongjian Li, Bohua Chen, Yang Yan, Hao Li, Bin Ling, Yonghong Tian, and Li Yuan. 2024. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model.

Amin Dada, Aokun Chen, C.A.I. Peng, Kaleb E. Smith, Ahmad Idrissi-Yaghir, Constantin Seibold, Jianning Li, Lars Heiliger, Xi Yang, Christoph M. Friedrich, Daniel Truhn, Jan Egger, Jiang Bian, Jens Kleesiek, and Yonghui Wu. 2023. On the impact of cross-domain data on german language models. In *Conference on Empirical Methods in Natural Language Processing*.

Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. 2019. Domain generalization via model-agnostic learning of semantic features. In *Neural Information Processing Systems*.

Xinjie Fan, Qifei Wang, Junjie Ke, Feng Yang, Boqing Gong, and Mingyuan Zhou. 2021. Adversarially adaptive normalization for single domain generalization. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 8208–8217.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. 2020. A brief review of domain adaptation.

Zainab Fatima, Shehnila Zardari, Muhammad Fahim, Maria Andleeb Siddiqui, Ag Asri Ag Ibrahim, Kashif Nisar, and Laviza Falak Naz. 2022. A novel approach for semantic extractive text summarization. *Applied Sciences*, 12(9):4479.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation.

Mark Glickman and Yi Zhang. 2024. Ai and generative ai for research discovery and summarization.

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

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re)warm your model?

Suchin Gururangan, Mike Lewis, Ari Holtzman, Noah A. Smith, and Luke Zettlemoyer. 2021. Demix layers: Disentangling domains for modular language modeling.

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.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *ArXiv*, abs/2305.14992.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

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.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023a. Continual pre-training of language models. In *The Eleventh International Conference on Learning Representations*.

Zixuan Ke, Yijia Shao, Haowei Lin, Hu Xu, Lei Shu, and Bing Liu. 2023b. Adapting a language model while preserving its general knowledge.

Blanka Frydrychova Klimova, Marcel Pikhart, Alice Delorme Benites, Caroline Lehr, and Christina Sanchez-Stockhammer. 2022. Neural machine translation in foreign language teaching and learning: a systematic review. *Education and Information Technologies*, 28:663–682.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *ArXiv*, abs/1810.09305.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. 2018. Deep domain generalization via conditional invariant adversarial networks. In *European Conference on Computer Vision*.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023. Large language models in finance: A survey.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, Chris White, Quanquan Gu, Jian Pei, and Liang Zhao. 2023. Domain specialization as the key to make large language models disruptive: A comprehensive survey.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. 2015. Learning transferable features with deep adaptation networks.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. 2022. Teaching language models to support answers with verified quotes.

Ning Miao, Yee Whye Teh, and Tom Rainforth. 2023. Selfcheck: Using llms to zero-shot check their own step-by-step reasoning. *ArXiv*, abs/2308.00436.

Le Thanh Nguyen-Meidine, Atif Belal, Madhu Kiran, Jose Dolz, Louis-Antoine Blais-Morin, and Eric Granger. 2020. Unsupervised multi-target domain adaptation through knowledge distillation.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. 2019. Moment matching for multi-source domain adaptation.

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. 2023. Progressive prompts: Continual learning for language models.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage?

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned language models are continual learners.

Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. 2022. Legal case document summarization: Extractive and abstractive methods and their evaluation.

Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun, and Jie Zhou. 2022. On transferability of prompt tuning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Kai Sun, Y. Xu, Hanwen Zha, Yue Liu, and Xinhsuai Dong. 2023. Head-to-tail: How knowledgeable are large language models (llm)? a.k.a. will llms replace knowledge graphs? *ArXiv*, abs/2308.10168.

Yashar Talebirad and Amirhossein Nadiri. 2023. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *ArXiv*, abs/2306.03314.

Arun James Thirunavukarasu, Refaat Hassan, Shathar Mahmood, Rohan Sanghera, Kara Barzangi, Mohanned El Mukashfi, and Sachin Shah. 2023a. Trialling a large language model (chatgpt) in general practice with the applied knowledge test: Observational study demonstrating opportunities and limitations in primary care. *JMIR Medical Education*, 9.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023b. Large language models in medicine. *Nature Medicine*, 29:1930–1940.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. 2023. Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera

Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2023. Clinical text summarization: Adapting large language models can outperform human experts.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. 2022. Spot: Better frozen model adaptation through soft prompt transfer.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. 2021. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35:8052–8072.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. 2022. Generalizing to unseen domains: A survey on domain generalization.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Hao Wen, Yuanchun Li, Guohong Liu, Shanhui Zhao, Tao Yu, Toby Jia-Jun Li, Shiqi Jiang, Yunhao Liu, Yaqin Zhang, and Yunxin Liu. 2023. Empowering llm to use smartphone for intelligent task automation. *ArXiv*, abs/2308.15272.

Ruoxi Xu, Yingfei Sun, Mengjie Ren, Shiguang Guo, Ruotong Pan, Hongyu Lin, Le Sun, and Xianpei Han. 2024. Ai for social science and social science of ai: A survey.

Hongbin Ye, Ningyu Zhang, Shumin Deng, Xiang Chen, Hui Chen, Feiyu Xiong, Xi Chen, and Huajun Chen. 2022. Ontology-enhanced prompt-tuning for few-shot learning. In *Proceedings of the ACM Web Conference 2022*, WWW '22. ACM.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021. AdaptSum: Towards low-resource domain adaptation for abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5892–5904, Online. Association for Computational Linguistics.

Liwen Zhang, Wei Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qi Qin, Yifei Li, Xingxian Liu, Zhiqiang Liu, Zhoufan Zhu, Anbo Wu, Xinnan Guo, and Yun Chen. 2023. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *ArXiv*, abs/2308.09975.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Xingxuan Zhang, Linjun Zhou, Renzhe Xu, Peng Cui, Zheyan Shen, and Haoxin Liu. 2022. Nico++: Towards better benchmarking for domain generalization. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16036–16047.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. 2023. A survey of large language models. *ArXiv*, abs/2303.18223.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2021a. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:4396–4415.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. 2021b. Domain generalization with mixstyle. *ArXiv*, abs/2104.02008.

A The training prompts

The prompt used to train the Llama2 model.

```
<s> [INST] «SYS» You are a helpful, respectful and honest assistant.  Always
answer as helpfully as possible while being safe.  Your answers should not
include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal
content.  Please ensure that your responses are socially unbiased and positive in
nature.

If a question does not make any sense, or is not factually coherent, explain
why instead of answering something not correct.  If you don't know the answer to
a question, please don't share false information.  «/SYS»

Summarize the following paragraph
<Dodument>
[/INST]
```

The prompt used to train the Bloom model.

```
<s>A chat between a curious user and an artificial intelligence assistant.  The
assistant gives helpful, detailed, and polite answers to the user's questions.
summarize the following paragraph <Document>
```

B Results of Bloom-1.1B and 3B

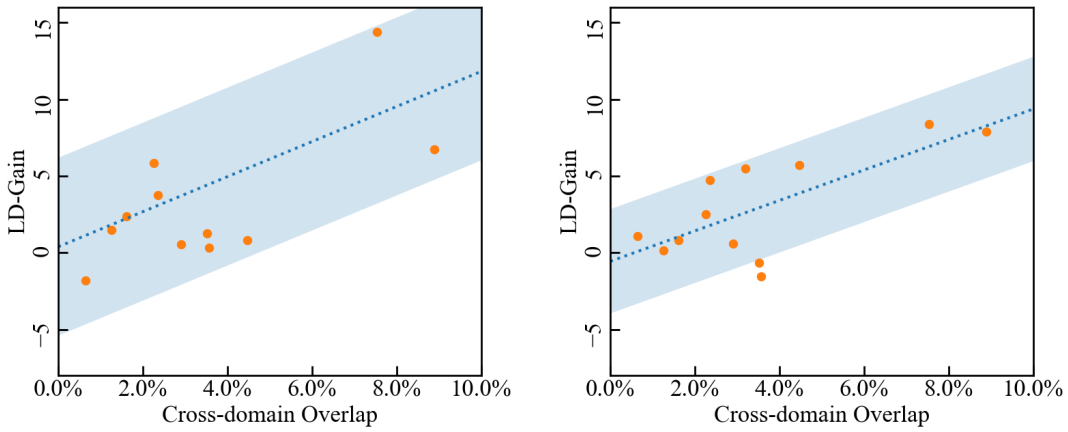
C Results of BERTScore

Training Set	Test Set	Compression Rate	Abstract Level	Learning Difficulty Coefficient	ROUGE Base	ROUGE Fine-tuned	ROUGE Improvement	Cross-domain Overlap	*LD-Gain*
PubMed						3.33	-0.39	2.35%	-102.12
SAMSum	CNNDM	12.95	20.22	261.85	3.72	6.13	2.41	8.89%	631.06
WikiHow						4.38	0.66	4.47%	172.82
CNNDM						3.87	-0.26	2.90%	-24.17
SAMSum	PubMed	7.08	13.13	92.96	4.13	4.78	0.65	3.51%	60.42
WikiHow						5.2	1.07	3.96%	99.47
CNNDM						1.32	-0.43	1.62%	-35.03
PubMed	SAMSum	4.86	16.76	81.45	1.75	2.04	0.29	0.64%	23.62
WikiHow						2.37	0.62	1.26%	50.50
CNNDM						2.43	-0.08	3.30%	-40.96
PubMed	WikiHow	13.05	39.23	511.95	2.51	2.66	0.15	2.25%	76.79
SAMSum						2.46	-0.05	7.53%	-25.6

Table 5: Results of single-domain adaptation on the Bloom-1.1B model.

Training Set	Test Set	Compression Rate	Abstract Level	Learning Difficulty Coefficient	ROUGE Base	ROUGE Fine-tuned	ROUGE Improvement	Cross-domain Overlap	*LD-Gain*
PubMed						7.38	1.03	2.35%	269.71
SAMSum	CNNDM	12.95	20.22	261.85	6.348	8.91	2.56	8.89%	670.34
WikiHow						5.61	-0.74	4.47%	-193.77
CNNDM						7.64	-0.96	2.90%	-89.24
SAMSum	PubMed	7.08	13.13	92.96	8.60	9.20	0.61	3.51%	56.71
WikiHow						6.66	-1.94	3.56%	-180.34
CNNDM						9.16	3.97	1.62%	323.36
PubMed	SAMSum	4.86	16.76	81.45	5.19	3.22	-1.97	0.64%	-160.46
WikiHow						0.66	-4.53	1.26%	-368.97
CNNDM						4.77	1.06	3.30%	542.67
PubMed	WikiHow	13.05	39.23	511.95	3.709	3.44	-0.27	2.25%	-138.23
SAMSum						5.00	1.30	7.53%	665.54

Table 6: Results of single-domain adaptation on the Bloom-3B model.



(a) Bloom-3B (b) Llama2-7B

Figure 5: Results of single-domain adaptation calculated by BERTScore.