AdaptEval: Evaluating Large Language Models on Domain Adaptation for Text Summarization

Anum Afzal
Technical University of Munich
anum.afzal@tum.de

Ribin Chalumattu
ETH Zürich
cribin@inf.ethz.ch

Florian Matthes
Technical University of Munich
matthes@tum.de

Laura Mascarell
ETH Zürich
lmascarell@inf.ethz.ch

Abstract

Despite the advances in the abstractive summarization task using Large Language Models (LLM), there is a lack of research that assess their abilities to easily adapt to different domains. We evaluate the domain adaptation abilities of a wide range of LLMs on the summarization task across various domains in both fine-tuning and in-context learning settings. We also present AdaptEval, the first domain adaptation evaluation suite. AdaptEval includes a domain benchmark and a set of metrics to facilitate the analysis of domain adaptation. Our results demonstrate that LLMs exhibit comparable performance in the in-context learning setting, regardless of their parameter scale.

1 Introduction

Large Language Models (LLM) have achieved remarkable improvements on a wide range of natural language processing tasks, including abstractive text summarization, the task of generating an abridged version of the most relevant information in a document (Basyal and Sanghvi, 2023). Recent works study the domain adaptation abilities of LLMs on the summarization task. However, the research is still limited to a single domain, such as news articles (Goyal et al., 2022; Zhang et al., 2023) or clinical reports (Van Veen et al., 2023). We argue that there is a lack of research across domains to better understand the abilities of these models to adapt to different targets.

In this paper, we assess the domain adaptation abilities of 11 models, including conventional encoder-decoder models and a wide range of LLMs in various parameter sizes, on the summarization task. In particular, we experiment with fine-tuning and in-context learning (ICL) settings and evaluate their performance across various domains (i.e. governmental, medical, and scientific), reporting scores on a collection of automatic—ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019)—and domain adaptation metrics. The latter includes domain vocabulary overlap (Yu et al., 2021), and our adaptations of G-eval (Liu et al., 2023) and token distribution shift (Lin et al., 2023) to the task.

The experimental results show the abilities of LLMs to adapt to the domain in the ICL setting. In particular, *small* models with 7b parameters achieve comparable performance to their larger counterparts with only two learning examples. However, G-eval highlights the difficulty of adapting to the medical domain. While the fine-tuned models achieve the best performance in terms of automatic scores, their adaptation to the domain vocabulary is inferior to the ICL setting. Finally, we release the domain benchmark and evaluation metrics as the first domain **Adapt**ation **Eval**uation suite (**AdaptEval**) to facilitate the evaluation of models and foster further research on this task.[1]

2 The Domain Adaptation Suite

2.1 Domains Benchmark

Our benchmark contains data from different datasets on the scientific, medical, and governmental domains. The final size of the domain datasets is listed in Table 1, after removing instances with extractive summaries, or extremely long summaries or sources as in Shaham et al. (2022).[2]

Science The data consists of scientific articles from the arXiv platform, where the human-written abstracts are used as reference summaries of the articles (Cohan et al., 2018).

Medical The medical domain comprises academic articles in the field of biomedical and life sciences from the PubMed dataset (Cohan et al., 2018). Similarly to arXiv, the article abstracts are regarded as abstractive summaries.

[1]AdaptEval code is available on AdaptEval.
[2]Deleted: 3% arXiv, 4% PubMed, and 0.4% GovReport.

Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U), pages 76–85

Domain	Train	Val.	Test
Science	203,037	6,436	6,440
Medical	119,924	6,633	6,658
Government	17,517	973	973

Table 1: Sizes of domain datasets.

Domain	Size	#W	#Sum W
Science	215,913	6,029.9	272.7
Medical	133,215	3,049.9	204.4
Government	19,466	9,409.4	553.4

Table 2: Total sizes of the domain datasets and average word count of source (#W) and summary (#Sum W).

Government The data comes from the GovReport dataset, a collection of reports on national policy issues paired with human-written executive summaries (Huang et al., 2021). The documents are 1.5 and 2.5 times longer than those from arXiv and PubMed, respectively.

2.2 Evaluation Metrics

The suite provides a set of metrics to evaluate the performance of summarization models and approaches across domains. Specifically, we include the standard summarization metrics ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2019), which measure n-gram and contextual similarity against a reference, respectively. To get better insights into their domain adaptation abilities, we also implement several metrics that assess the domain language. We describe them in the rest of the section.

Domain Vocabulary Overlap (DVO) We compute the percentage of domain vocabulary in the generated output as in Yu et al. (2021). The domain vocabulary consists of the top 10k most frequent words in the domain excluding stopwords.

Domain Token Distribution Shift Lin et al. (2023) analyzes the impact of LLM alignment and proposes to measure the token distribution shifts between base models and their aligned counterparts. We adopt the token distribution shift approach to domain adaptation. Specifically, we focus on the domain vocabulary (i.e. 10k most frequent words) and analyze the effects of adaptation strategies, such as ICL and fine-tuning on their distribution.

Formally, given a prompt p, we first use the fine-tuned model to generate a summary by greedy decoding, where the summary is represented as a sequence of tokens $S = \{s_0, ..., s_T\}$ from the model vocabulary $\mathcal{V}$, such that $s_t \in \mathcal{V}$ for $0 < t < T$. Next, we process each token in S sequentially. At each step t, we get the probability distribution of the next token prediction given p and the prior context $p(\cdot \mid \mathbf{s}_{<t}, \mathbf{p})$ using both fine-tuned and base models. In the in-context learning setting, we use the same model, but the adapted approach extends the prompt p with learning examples.[3] Finally, we rank the tokens in both distributions according to their probability and provide *KL-divergence* scores and the *token shift rate* of those tokens in the vocabulary domain. While the former represents their distribution similarity, the latter computes the frequency at which the adapted approach predicts a token from the vocabulary domain that is not among the top three predictions of the base model.

Reference-free evaluation with GPT-4 G-eval uses GPT-4 (OpenAI, 2023) with chain-of-thought prompting (Wei et al., 2022) to evaluate summaries across quality features, such as coherence or fluency, achieving high correlation with human judgments (Liu et al., 2023). Similarly, we design a prompt to score the degree to which a summary adheres to the domain language on a scale from 1 to 5. Our prompt includes the reasoning steps generated by GTP-4 as in Liu et al. (2023) (see Appendix B).

3 Domain Adaptation Task

We assess the performance of 11 models across domains in both fine-tuning[4] and ICL settings.

3.1 Models Selection

We select a wide variety of models from the conventional encoder-decoder transformer models—BART (Lewis et al., 2020) and PEGASUS-X (Phang et al., 2022)—to the recent instruction-based LLMs. The latter includes open-source models from the Llama2 family (Touvron et al., 2023), Vicuna (Chiang et al., 2023), Falcon (Almazrouei et al., 2023), and Mistral AI (Jiang et al., 2023). For each model family, we consider various model sizes ranging from 7b to 70b parameters, if available. Additionally, we consider the close-source model ChatGPT from OpenAI. We provide the checkpoints and technical details in Appendix A.

[3]The method can also be applied to compare models of different parameter scales in different adaptation settings.

[4]We exclude GovReport from fine-tuning on 5k and 10k samples, since the train set doesn't have enough documents to fit into the models context window of 4096 tokens—only 1148 instances with maximum 4k length in the training split.

	Medical			Science			Government		
	BERTScore	DVO	ROUGE	BERTScore	DVO	ROUGE	BERTScore	DVO	ROUGE
Zero-shot Setting									
PEGASUS-X	0.690	6.28	3.55	0.538	11.98	5.85	0.736	5.58	9.06
Falcon 7b	0.811	31.87	13.68	0.810	30.16	14.54	0.821	31.49	13.86
Llama2 7b	0.783	21.15	10.94	0.818	28.61	18.33	0.845	34.36	18.86
Mistral 7b	0.788	24.78	9.44	0.806	28.81	13.68	0.815	31.18	12.02
Vicuna 7b	0.727	9.49	2.11	0.781	23.94	7.93	0.813	30.69	10.80
Llama2 13b	0.764	20.78	6.26	0.783	23.48	8.58	0.797	24.04	10.80
Vicuna 13b	0.745	15.76	1.58	0.763	19.07	4.43	0.783	27.18	7.17
Falcon 40b	0.816	35.51	13.85	0.822	34.98	17.59	0.827	35.51	13.85
Llama2 70b	*0.842*	35.50	*24.59*	*0.837*	35.22	23.35	*0.855*	36.05	21.48
ChatGPT	**0.844**	36.69	**24.81**	**0.838**	36.58	**23.95**	**0.859**	37.73	**22.34**
GPT-4o mini	0.843	**41.04**	22.26	0.834	**40.85**	20.16	0.856	**41.51**	21.12
Two-shot Setting									
Llama2 7b	0.819	35.95	21.11	0.824	35.34	20.92	0.847	30.22	17.39
Mistral 7b	*0.816*	32.05	21.30	0.802	23.61	17.76	0.844	30.08	**19.21**
Vicuna 7b	0.831	36.29	21.54	0.827	34.65	20.31	0.851	30.28	17.29
Llama2 13b	*0.820*	35.02	19.00	0.809	32.30	18.97	0.814	29.92	14.30
Vicuna 13b	0.822	35.51	19.69	*0.807*	33.32	14.86	0.789	29.34	8.34
Llama2 70b	**0.845**	37.61	*22.40*	**0.842**	36.65	23.03	0.851	29.59	18.72
ChatGPT	0.841	**38.58**	22.92	0.837	**38.39**	**23.15**	**0.853**	**30.44**	16.82
GPT-4o mini	0.842	30.64	**23.18**	0.835	29.14	21.47	0.850	30.40	16.04
Fine-tuning Setting									
BART	0.852	**37.03**	24.80	0.844	*34.15*	22.20	0.856	25.14	28.44
PEGASUS-X	0.850	28.72	**31.18**	0.852	**34.61**	**28.11**	**0.868**	22.07	**31.98**
Llama2 7b[1]	0.859	33.61	25.81	0.858	33.06	25.30	0.850	29.30	24.81
Llama2 7b[2]	0.861	35.15	26.00	0.856	30.49	25.46	x	x	x
Llama2 7b[3]	*0.862*	33.71	26.81	0.854	27.43	25.35	x	x	x
Mistral 7b[2]	**0.863**	35.81	27.17	**0.863**	34.00	27.29	0.833	21.66	23.08
Llama2 13b[2]	*0.862*	*35.28*	26.26	*0.860*	32.67	26.47	x	x	x

Table 3: BERTScore F_1, DVO (%), and the geometric mean of ROUGE-1/2/L (ROUGE) of all models across the three domains. The value 'x' implies that the model was not evaluated under those settings. $^1/^2/^3$ indicate fine-tuning with 1k, 5k, and 10k instances, respectively.

3.2 Results

Table 3 shows the performance of the models across domains in terms of ROUGE, BertScores, and DVO. We observe that the model size has a direct impact on their overall performance in the zero-shot setting; however, this performance gap is considerable reduced in the ICL setting with only two learning examples. In fact, the scores of the small 7b models are comparable to the large Llama 70b or the even larger ChatGPT. To validate these results, we compute the token distribution shift between models of different sizes in the two-shot setting (Table 4). The scores reflect that their probability distributions are very similar, confirming that there are no major differences in their performance.

In contrast, the fine-tuning results in Table 3 are mixed. Overall, the models outperform their counterparts in the two-shot setting in terms of ROUGE scores; however, there is a decrease in DVO. In particular, PEGASUS-X achieves the best ROUGE scores. We argue that this is attributed to the model's fine-tuning process, since the parameters are adjusted to optimize on ROUGE. Additionally, BART achieves the highest DVO despite its small parameter size (110M). Johner et al. (2021) point out to the model's tendency to generate highly extractive summaries, which favours the use of domain vocabulary. Finally, the token shift rate and KL-divergence scores between the base and fine-tuned models are higher than in the two-shot setting. However, we observe that most distribution shifts are due to stylistic tokens, as also reported in Lin et al. (2023) between the base and their aligned LLMs.

To confirm these findings, we also evaluate the summaries using GPT-4 shown in Table 5, which have a strong correlation with human judgments, along with our addition to measure domain adaptation, on a random sample of 25 articles.[5] The

[5]Due to the costs of using GPT-4 with large prompts, we

		Science		Medical		Government	
base	*2-shot*	KL	TSR	KL	TSR	KL	TSR
Llama2 7b	vs. 7b	19.70	92.14	19.27	97.44	17.40	94.33
Mistral 7b	vs. 7b	13.88	91.33	14.01	95.40	13.40	90.00
Vicuna 7b	vs. 7b	17.67	92.35	18.32	93.89	15.42	94.04
Llama2 13b	vs. 13b	15.58	96.95	16.53	96.76	14.67	98.82
Vicuna 13b	vs. 13b	18.12	97.13	17.34	90.70	16.79	99.10
Llama2 70b	vs. 70b	16.78	95.68	17.12	98.19	13.10	92.36
2-shot	*2-shot*	KL	TSR	KL	TSR	KL	TSR
Llama2 13b	vs. 7b	0.21	2.87	0.38	1.67	0.32	10.38
Vicuna 13b	vs. 7b	0.25	2.07	0.38	4.57	0.24	0.00
Llama2 70b	vs. 13b	0.47	5.18	0.31	3.50	0.49	4.92
Llama2 70b	vs. 7b	0.43	3.92	0.46	5.01	0.54	6.88
base	*FT*	KL	TSR	KL	TSR	KL	TSR
Llama2 7b	vs. 7b	0.81	12.40	0.35	4.70	21.49	15.15
Mistral 7b	vs. 7b	0.52	11.54	0.37	4.42	0.18	3.21
Llama2 13b	vs. 13b	0.51	6.84	0.48	7.32	x	x

Table 4: Effect of different model sizes, two-shot in-context learning, and Fine-Tuning in terms of token distribution shift scores—KL divergence and Token Shift Rate (%) calculated over 10 samples. Two-shot has the major impact on the models' predictions. The low scores between different model sizes indicate that parameter size does not have a significant effect on domain adaptation in the two-shot setting.

scores on arXiv data are consistent with our previous results, showing that ICL achieves the best performance, and the model parameter size does not have a significant impact. However, PubMed obtains remarkably low scores, which highlights the difficulty of the models to adapt to the medical domain. The LLMs however, find it easier to adapt to the Government domain.

3.3 Manual Evaluation

Two in-house domain experts perform a blind manual evaluation of the same arXiv samples used in GPT-4 evaluation (Table 5). The setting comprises of 25 random arXiv articles paired with four different summaries generated with Llama2 (7b and 70b) in the two-shot setting, fine-tuned Llama2 (7b) and PEGASUS-X. To avoid biases, we randomly shuffle the evaluation instances and their summaries for each annotator.

We ask the annotators to rank the generated summaries according to how well the vocabulary and style of the outputs adapt to the scientific domain. The task is especially challenging when the summaries contain similar vocabulary. Therefore, we focus on the relative performance of the models; that is, their agreement on an output being ranked higher than the other. The final Cohen's κ inter-annotator agreement is 0.4. The results show that the annotators consistently rated the outputs of both Llama2 7b and 70b in the two-shot scenario among the top two positions of the ranking—60% and 52%, respectively—whereas the fine-tuned models were the least preferred—only 12% (Llama2 7b) and 16% (Pegasus-X) rated on top.

only report the scores on four models outputs of 25 random instances.

4 Related Work

Some recent works evaluate the domain adaptation abilities of LLMs on the summarization task, albeit limited to a specific domain. Van Veen et al. (2023) focus on clinical data and tackle the summarization of electronic health records. They evaluate eight different LLMs across six datasets in the same domain. Fu et al. (2024) investigate whether model size has an impact on the summarization performance of business meeting transcripts. The results show that smaller LLMs cannot outperform their larger counterparts (from 7b to 70b parameters), even after fine-tuning, except for FLAN-T5 with 780M parameters (Chung et al., 2022). In contrast, Zhang et al. (2023) provides a benchmark for text summarization of news articles and concludes that instruct-tuning rather than model size is the key to text summarization with LLMs. Similarly, Goyal et al. (2022) propose also a news summarization benchmark and compare the performance between conventional encoder-decoder and instruction-based models. Prior to the LLM era, Yu et al. (2021) explored domain adaptation

2-shot	DA (ours) arXiv	DA (ours) PubMed	DA (ours) GovReport	Coherence arXiv	Coherence PubMed	Coherence GovReport	Fluency arXiv	Fluency PubMed	Fluency GovReport
Llama2 7b	4.20	1.0	4.04	3.80	2.0	3.96	2.72	2.0	2.96
Llama2 70b	3.96	1.0	4.40	3.20	1.0	3.96	2.56	1.0	3.00
FT									
Llama2 7b	3.48	2.0	4.16	2.08	2.0	3.40	2.04	2.0	2.84
PEGASUS-X	3.88	2.8	4.40	2.88	2.0	3.72	2.40	2.0	2.72

Table 5: Evaluation scores using GPT-4 on 25 random samples from the arXiv, PubMed and GovReport datasets in terms of coherence (1-5), fluency (1-3), and our Domain Adaptation (DA) (1-5).

techniques in a low-resource setting, such as fine-tuning and second pre-training of encoder-decoder summarization models on a wide range of datasets.

5 Conclusion

We evaluate the domain adaptation abilities of Large Language Models across scientific, medical, and governmental domains using a set of adapted evaluation metrics. Additionally, we release AdaptEval, an evaluation suite that facilitates the analysis of domain adaptation. Our experiments show that smaller LLMs exhibit domain-shift challenges, but they are able to achieve comparable performance to larger LLMs when provided with only two learning examples. In contrast, fine-tuning does not have a significant impact on the vocabulary domain, but only on stylistic tokens. Overall, the G-eval scores indicate that the medical domain is challenging for these models. We expect our work to encourage and facilitate further research on domain adaptation with LLMs across domains. We plan to continue this research in future work.

Limitations

To fairly compare the performance of the different models, we generally restricted our evaluation to those models with context window of 4096. An exception is the language model BART with a context window of 1024. Additionally, due to the high costs of performing human evaluations on multiple domains, we only annotated ArXiv data to reaffirm the results obtained through the automatic metrics. Our goal is to facilitate the evaluation of models across domains to the research community. Therefore, our suite consists of a set of metrics to evaluate domain adaptation and general summarization quality, allowing for a comprehensive comparison of the models performance on multiple datasets. Lastly, given the cost associated with GPT-4, we performed LLM-based evaluation on only 25 random samples.

Ethics Statement

Throughout our experiments, we strictly adhere to the ACL Code of Ethics. Since we used already established open-source benchmark datasets, the concern of privacy does not apply. The manual evaluation was performed by in-house domain experts, who receive a full salary. They were informed about the task and usability of data in the research. Their annotations were stored anonymously, mitigating any privacy concerns. Through our fine-tuning strategies, no additional bias was introduced into the models, other than what might already be part of the model weights or the benchmark dataset. The goal of the research is to evaluate the domain adaptation capabilities of existing models on a text summarization task. The results and discussions in this paper are meant to further promote research in the area of domain-specific language modeling with an over-arching goal of bridging the gap between academia and application. All training scripts and trained models will be made available to the research community.

Acknowledgements

This project is supported by Ringier, TX Group, NZZ, SRG, VSM, viscom, the ETH Zurich Foundation, and the German Federal Ministry of Education and Research (BMBF) grant 01IS17049 Software Campus 2.0 (TU München).

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023.

The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

Lochan Basyal and Mihir Sanghvi. 2023. Text summarization using large language models: A comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models. *arXiv preprint arXiv:2310.10449*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv: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.

Xue-Yong Fu, Md Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen, and Shashi Bhushan TN. 2024. Tiny titans: Can smaller large language models punch above their weight in the real world for meeting summarization? *arXiv preprint arXiv:2402.00841*.

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

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Timo Johner, Abhik Jana, and Chris Biemann. 2021. Error analysis of using BART for multi-document summarization: A study for English and German language. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 391–397, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

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

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. *arXiv preprint arXiv:2312.01552*.

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, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jason Phang, Yao Zhao, and Peter J Liu. 2022. Investigating efficiently extending transformers for long input summarization. *arXiv preprint arXiv:2208.04347*.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. SCROLLS: Standardized CompaRison over long language sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12007–12021, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, William Collins, Neera Ahuja, et al. 2023. Clinical text summarization: Adapting large language models can outperform human experts. *arXiv preprint arXiv:2309.07430*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In

Advances in Neural Information Processing Systems, volume 35, pages 24824–24837. Curran Associates, Inc.

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.

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.

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

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, L. Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *ArXiv*, abs/2305.11206.

A Technical Details

The fine-tuning and inference procedure was done by leveraging Nvidia A100-80GB GPUs.

A.1 Zero-shot Setting

We used the instruct-tuned or chat versions of the models. As for ChatGPT, we used the OpenAI API[6] and the latest snapshot available, `gpt-3.5-turbo-0613` from June 13th, 2023. For zero-shot setting, we used Llama2 (7b)[7], Llama2 (13b)[8], Llama2 (70b) [9], Vicuna (7b)[10], Vicuna (13b)[11], Falcon (7b)[12], Falcon (40b)[13], and Mistral AI (7b)[14].

When generating summaries, we sample a maximum of 256 tokens for the arXiv and PubMed datasets, while scaling to 1024 tokens for the GovReport dataset, as is standard procedure in other contemporary publications. The prompts used 0-shot and 2-shot settings for generating the summaries is shown in Table 7.

A.2 In-context Learning Setting

We used the same model checkpoints as the ones from zero-shot settings for in-context learning. We excluded Falcon from in-context learning, since its context window of 2048 is too small to fit 2 learning examples.

A.3 Fine-tuning Setting

The links to all fine-tuned models is displayed in Table 6.

Language Models We used HuggingFace Transformers (Wolf et al., 2020) and Microsoft Deepspeed library for distributed training.[15] We fine-tuned BART[16] and PEGASUS-X[17] on the training split and a context window of 1024 and 4096, respectively. All models were fine-tuned for 4 epochs with a learning rate of $8e-4$ and batch size of 64.

[6] https://platform.openai.com/

[7] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf/

[8] https://huggingface.co/meta-llama/Llama-2-13b-chat-hf/

[9] https://huggingface.co/meta-llama/Llama-2-70b-chat-hf/

[10] https://huggingface.co/lmsys/vicuna-7b-v1.5

[11] https://huggingface.co/lmsys/vicuna-13b-v1.5

[12] https://huggingface.co/tiiuae/falcon-7b

[13] https://huggingface.co/tiiuae/falcon-40b

[14] https://mistralai/Mistral-7B-Instruct-v0.1

[15] https://github.com/microsoft/DeepSpeed

[16] https://huggingface.co/facebook/bart-base

[17] https://huggingface.co/google/pegasus-x-large

	Science	Medical	Government
BART	bart-arxiv-1024	bart-pubmed-1024	bart-govreport-1024
PEGASUS-X	bigbird-pegasus-arxiv-4096	bigbird-pegasus-pubmed-4096	bigbird-pegasus-govreport-4096
Llama2 7b[1]	Llama-2-7b-arxiv-4096	Llama-2-7b-pubmed-4096	Llama-2-7b-govreport-4096
Llama2 7b[2]	Llama-2-7b-arxiv-4096	Llama-2-7b-pubmed-4096	x
Llama2 7b[3]	Llama-2-7b-arxiv-4096	Llama-2-7b-hf-pubmed-4096	x
Llama2 13b[2]	Llama-2-13b-arxiv-4096	Llama-2-13b-pubmed-4096	x
Mistral 7b[2]	Mistral-7B–arxiv-4096	Mistral-7B-pubmed-4096	Mistral-7B-govreport-4096

Table 6: Links to all fine-tuned models repositories. The value 'x' implies that the model was not evaluated under those settings. [1]/[2]/[3] indicate fine-tuning with 1k, 5k, and 10k instances, respectively.

Large Language Models We included Llama2 (7b)[18], Llama2 (13b)[19], and Mistral AI [20] for LLM fine-tuning. We fine-tuned the models for 1 epoch using the HuggingFace Trainer API and LoRA on a training subset consisting of samples with a maximum length of 4096, such that they can fit in the context window without truncation. Since Zhou et al. (2023) argue that 1k samples are enough to fine-tune LLMs, we experimented with 1k, 5k, and 10k training samples. Since models do not show any performance increase when trained on more than 5k samples, we opted to train on Llama2 (13b) and Mistral AI on 5k samples. We selected the LoRA parameters r=64, alpha=16, and a dropout of 0.1. Furthermore, we used the paged AdamW optimizer with a beta2 value of 0.999 and a learning rate of $2e-4$ with a constant learning rate strategy. We did not fine-tune Vicuna, since we only used the non-instruction tuned models in this setting. We excluded Falcon from fine-tuning as it only supports a context window of 2048, and therefore, it cannot be fairly compared against the other models with a context window of 4096.

B LLM Prompting

Table 7 and Table 8 illustrate the prompts used to generate summaries and to score the domain adaptation of summaries using GPT-4, respectively. For evaluation, we use the prompts introduced by Liu et al. (2023) for Coherence and Fluency. However, we craft our own prompt that asseses model's ability to adapt to a new domain by evaluating the generated summaries.

[18] https://huggingface.co/meta-llama/Llama-2-7b

[19] https://huggingface.co/meta-llama/Llama-2-13b

[20] https://huggingface.co/mistralai/Mistral-7B-v0.1

C Sample Summaries

Table 9 shows the summaries generated by Llama2 7b under zero-shot, two-shot and fine-tuning setting.

0-SHOT PROMPT

You are an expert at summarization. Proceed to summarize the following text.
TEXT: {article}
SUMMARY:

FEW-SHOT PROMPT

You are an expert at summarization. Proceed to summarize the following text.
TEXT: {article}
SUMMARY: {summary}
Proceed to summarize the following text.
TEXT: {article}
SUMMARY: {summary}
...
TEXT: {article}
SUMMARY:

Table 7: The prompt in the Benchmark for generation of domain-specific summaries using Large Language Models.

SYSTEM PROMPT

You will be given one summary written for an article. Your task is to rate the summary on one metric. Please make sure you read and understand these instructions very carefully. Please keep this document open while reviewing, and refer to it as needed.
Evaluation Criteria: {criteria}
Evaluation Steps: {steps}

USER PROMPT

Example:
Source Text: {document}
Summary: {summary}
Evaluation Form: Please provide your response in two parts. First the evaluation score for the defined metrics, followed by an explanation for the score. Please limit your response to 30 words.
{metric_name}

DOMAIN ADAPTATION SCORE CRITERIA

Domain Adaptation(1-5) - the degree to which the summary adheres to the domain-specific language. A good summary employs domain-specific terminology and conveys the sense that model comprehends and encapsulates domain-specific knowledge. It resembles the content that would authored by a domain expert. Annotators were also asked to penalize summaries that didn't adhere to domain-specific knowledge, and rather used simple words.

DOMAIN ADAPTATION SCORE STEPS

1. Read the article carefully and understand the domain it belongs to.
2. Read the summary and check if it contains domain-specific terminologies and concepts, and if it is able to concisely summaries the domain specific concept in the article.
3. Assign a score for domain adaptation based on the Evaluation Criteria.

Table 8: The Domain Adaptation (DA) prompt in the Benchmark to evaluate the domain adaptation capabilities of the models using a GPT-4.

Zero-shot Llama2 7b
mirnas are small non coding rnas that play important roles on the onset and progression of cancer . distinct mirnas regulate various cellular processes including cell cycle regulation , apoptosis , migration and invasion . altered expression of mirnas has been observed in breast cancer and is believed to play important roles including cell proliferation regulation , cell survival and tumorigenesis . furthermore , mirnas have been shown to play a critical role in cancer stem cell maintenance and tumorigenesis . emerging evidence suggests that mirnas may also be a attractive therapeutic target for cancer . understanding the specific functionalities of distinct mirnas and their role in disease may provide a new approach for cancer diagnosis and potentially new therapeutic strategies. reference listing: lee , j.a. (2011). mi rna : a duel role player in breast cancer progression . breast cancer : a global view . springer . 10.10079783642-135704_19 . 19 . 277 286 . abstract: micro rnas

Two-shot Llama2 7b
Small non-coding RNAs, such as siRNA, miRNA, piRNA, and snRNA, are transcribed into mRNA but remain untranslated in eukaryotic cells. They play a crucial role in regulating gene expression and have been implicated in the onset and progression of tumorigenesis. Emerging evidence suggests that small non-coding RNAs may represent a novel cancer therapy. Unlike mRNA, small non-coding RNAs are transcribed but never translated. Some small non-coding RNAs are transcribed from non-coding regions between genes, while others are transcribed together with coding mRNAs from the coding region of the genome. Altered expression of small non-coding RNAs has been observed in breast cancer, resulting in aberrant gene expression. Studies have shown that a high frequency (72.8%) of gene copy number abnormality in small non-coding RNA-containing regions occurs in human breast cancer. Decreased expression of certain small non-coding RNAs, such as mir-17/20, has been

Fine-tuned Llama2 7b
mirnas are small non - coding rnas that regulate gene expression and tumorigenesis . mirnas are involved in the regulation of the cell cycle , cellular proliferation , metastasis and drug resistance of breast cancer . the expression of mirnas is dysregulated in breast cancer . the dysregulated mirnas may serve as novel biomarkers for breast cancer . mirnas may serve as targets for gene therapy either alone or as an adjuvant to conventional therapy .

Table 9: Summaries generated by Llama2 7b under zero-shot, two-shot and fine-tuning setting for a sample article (id = 2) from PubMed test set.