Enhancing Biomedical Text Summarization and Question-Answering: On the Utility of Domain-Specific Pre-Training

University of Technology Sydney participation in BioASQ Task 11b Phase B

Dima **Galat**, Marian-Andrei **Rizoiu**

University of Technology Sydney (UTS), Australia

Abstract
Biomedical summarization requires large datasets to train for text generation. We show that while transfer learning offers a viable option for addressing this challenge, an in-domain pre-training does not always offer advantages in a BioASQ summarization task. We identify a suitable model architecture and use it to show a benefit of a general-domain pre-training followed by a task-specific fine-tuning in the context of a BioASQ summarization task, leading to a novel three-step fine-tuning approach that works with only a thousand in-domain examples. Our results indicate that a Large Language Model without domain-specific pre-training can have a significant edge in some domain-specific biomedical text generation tasks.

Keywords
natural language processing, biomedical summarization, biomedical question answering, transfer learning, language modeling, domain-specific pre-training, BioASQ, CEUR-WS

1. Introduction

The fields of question-answering and summarization have witnessed significant advancements in recent years, with a shift from classification-based extractive approaches to the emergence of abstractive summarization models. This transition has been driven by the superior performance and enhanced generalization capabilities exhibited by abstractive models, effectively blurring the boundary between long-form question answering and summarization. This paper addresses the summarization challenge presented by BioASQ Task B Phase B in the biomedical domain, for which we propose a novel approach.

The healthcare sector holds immense potential for leveraging health research data sharing to enhance clinical care, informed decision-making, and scientific discovery [1]. Sharing biomedical and healthcare studies and research data with the wider public requires robust and efficient methods. *Large pre-trained language models* (LLMs) have emerged as promising candidates for this purpose. LLMs have the potential to store medical knowledge while accommodating variations in data and application tasks [2]. This paper aims to analyze the impact of the training
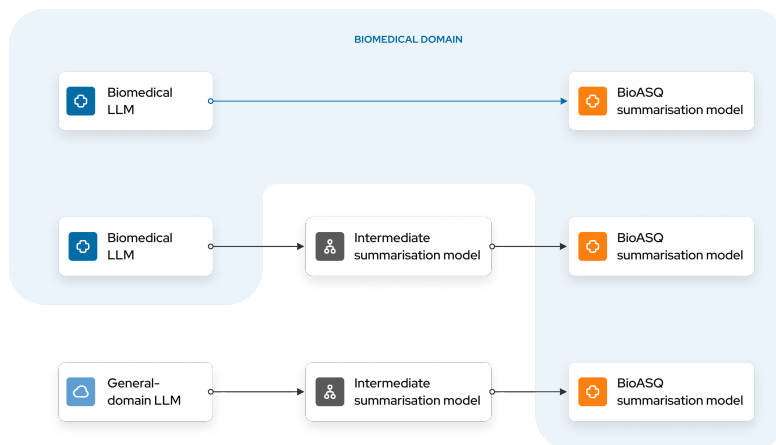
Figure 1: Schematic representation of LLM usage in practice. The blue arrow in the top row represents the two-phase (training and then fine-tuning) transfer learning approach typically seen in literature. The middle and bottom rows represent the task-specific fine-tuning approach proposed in this paper. Our results suggest that in-domain pre-training does not improve the BioASQ summarization performance even when summarization training is introduced in the middle (second row). In this diagram, the orange area indicates in-domain training.

process on LLMs' ability to store biomedical knowledge, explicitly focusing on their utilization for a question-answering and summarization task.

Traditionally, achieving state-of-the-art performance on natural language processing tasks involves a two-phase approach [3, 4] that is shown in blue in the top row of Fig. 1: pre-training the models on an extensive range of texts and topics, followed by task-specific fine-tuning [4, 5, 2]. This approach has revolutionized various areas of natural language processing [6, 7, 4], with LLMs such as BERT, GPT, and BART demonstrating remarkable capabilities. However, pre-training models is a time-consuming and a resource-intensive process, and the literature lacks comprehensive insights into the performance of these models for domain-specific applications with limited data availability. Therefore, this study aims to address this gap by examining the performance of LLMs in the context of the BioASQ summarization task.

This paper investigates two open questions concerning biomedical domain question-answering and text summarization tasks. Over the past five years, the biomedical domain has increasingly relied on in-domain pre-training and fine-tuning of BERT [4] for a wide range of datasets and benchmarks [8, 9, 10, 11, 12]. In-domain pre-training has proven effective in enhancing performance for discriminatory biomedical tasks. However, BERT's architecture is not optimized for text generation tasks [13, 14], lacking an autoregressive decoder to generate tokens based on previously generated ones. Consequently, BERT is suboptimal for generation tasks, necessitating exploring alternative approaches. Previous studies evaluating biomedical models across diverse tasks have not reported results on generation problems due to using non-autoregressive models [10]. The first question is **is there a better-suited architecture for biomedical text generation tasks?** A significant amount of research suggests that domain-specific pre-training significantly outperforms mixed-domain pre-training. However, we could

not find any convincing evidence for supporting this belief when it comes to text generation problems [8, 9, 10, 11, 15, 16, 12]. The second question is **do LLMs need to be pre-trained in domain to achieve optimal performance?**

We answer the above two questions. To investigate the efficacy of domain-specific pre-training and fine-tuning for biomedical text generation, we propose an alternative three-step approach (shown in the bottom row of Fig. 1). In this approach, we initially train a general-domain LLM, followed by fine-tuning for a specific task in the general domain (text summarization) and subsequent fine-tuning for the target biomedical domain task. Contrary to established theories in the biomedical domain [10, 15, 16], our findings suggest that having a large task-specific dataset can be more valuable than domain-specific pre-training for biomedical text generation tasks. This approach aligns with studies indicating that diverse pre-training objectives, larger and more diverse datasets, and tasks contribute to the robustness of the fine-tuning process even without domain adaptation [14, 17].

We explore alternative architectures for biomedical text generation. In this study, we focus on BART [5], a comprehensive architecture that incorporates pre-training objectives from both BERT [4] and GPT [18] models. BART has demonstrated state-of-the-art performance in abstractive dialogue, question-answering, and summarization tasks, making it particularly effective for text generation and comprehension. Our experimental results showcase the benefits and effectiveness of utilizing the BART architecture for transfer learning techniques in a context of a biomedical summarization task.

The main contributions of this work can be summarized as follows:

- Evaluating the advantages of domain-specific pre-training in the context of text generation tasks.
- Evaluating the impact of task-specific training on improving text generation tasks.
- Assessing the performance of BART, an encoder with an auto-regressive decoder architecture, in the biomedical question answering task B of BioASQ 11.

2. Related work

We are looking for a LLM which has an architecture suitable for long-form question answering and has been trained on relevant in-domain data. There are several important model architectures and pre-training objectives used to optimize the models worth considering [18, 4, 5].

First, lets briefly mention BERT [4] in the context of text generation, since most biomedical Transformer-based [3] models still rely on this architecture. BERT does not have an autoregressive decoder, preventing it from generating text. Despite this fact, a well-known summarisation approach called PreSumm [13] uses this architecture by inserting additional tokens for teaching models which sentences should be included in the summary. We followed the process proposed by the authors while using a BioBERT [9] model; we first trained an extractive summariser, which did perform a little better on BioASQ data than a regular BERT trained the same way. Unfortunately, when training an abstractive summarization architecture, PreSumm [13] process uses a randomly initialised Transformer [3] for a decoder. It appears that there is a significant mismatch between this decoder and a BioBERT [9] encoder leading to unstable abstractive

fine-tuning process and poor generation outputs in our experiments. Based on these findings, we have concluded that BERT is a wrong architecture to be using for text generation tasks.

BART [5] is an architecture that uses an encoder with an auto-regressive decoder, similarly to the original Transformer [3]. BART relies on an architecture which can be seen as generalising BERT (because it also uses a bi-directional encoder) and GPT [19] (because it also uses the left-to-right decoder). This model is using a masked language modeling objective (also known as denoising) introduced by BERT [4] and adds two additional denoising objectives (token deletion and sentence permutation). Authors conduct experiments that are focused on text generation, and show that denoising objectives are particularly well-suited for summarization tasks. Because it can be easily fine-tuned directly for generation tasks, authors achieved a remarkable success on a wide range of abstractive summarization and long-form question answering problems [5].

BioBART [15] is a BART model pre-trained on PubMed [20] abstracts. Authors have reported that they have trained without one of the objectives proposed by BART, namely the sentence permutation, showing that models trained without this objective have a better performance. Overall, this is the only study that we are aware of that applies a LLM to a range of generation tasks and reports the results (another BioGPT [16] study we found has not reported any numeric results on text generation problems). We are also not completely convinced that some of the results, like those reported for a BioASQ task could not be a result of a random chance, since the differences in the scores are very small and there are a few possible sources of non-determinism in training and generation procedures we discuss later in this paper.

3. Our Contribution

In the biomedical domain, the majority of models we have reviewed are focused on the pre-training process, perhaps because pre-training data is readily available [8, 9, 10, 11, 15, 16, 12]. However, question answering and summarization are plagued by a lack of a large domain specific dataset for fine-tuning LLMs directly for text generation problems. More specifically, when we are looking at the biomedical text generation tasks, it's hard to find a large (and clean) sequence-to-sequence dataset for fine-tuning for a long-form question answering and summarization. BioASQ is the closest dataset currently available, however it is still a few orders of magnitude away from what we would require to fine-tune a LLM for a previously unseen generation task. Therefore, we conclude that this two-step fine-tuning process offers a limited utility for this problem.

Following a conventional transfer learning definition we use a task to refer to training on labeled data, seeking to transfer the knowledge from a source task and a source domain ($\mathcal{T}_S$ and $\mathcal{D}_S$) to a target task and a target domain ($\mathcal{T}_T$ and $\mathcal{D}_T$) [21, 22]. One of the common transfer learning scenarios involves learning the tasks sequentially, one after another; and we could also have an intermediate fine-tuning task making it a three-step fine tuning process, where a second step is only used to get a representation that is more suitable for the task of summarization in a biomedical domain. This means that an intermediate $\mathcal{T}_{inter}$ (which could be both in/out domain) should lead to a performance improvement in $\mathcal{T}_T$. This could be potentially useful, since task-domain specific data is hard to come by.

Since we need to perform text generation, a reasonable option is to train for an $\mathcal{T}_{inter}$ which teaches the model to perform this task. Unfortunately, large question answering and summarization datasets like SQUAD [23] and CNN/DM [24] have nothing to do with biomedical domain, but because we need 10-100 times more biomedical summarization data than what we have available, we believe that task-specific datasets could offer just as much value as a domain-specific pre-training. We believe that CNN/DM is the most suitable (clean, large, easily available) task-specific dataset; especially because summaries there are typically closely related to source sentences, which is also the case with the BioASQ data. Moreover, lengths of summaries are similar to those in BioASQ. Therefore, we are interested in this task, even though a newsmedia domain would likely have completely different marginal probability distributions of generated text. This approach means that in addition to sequential transfer learning (two and three step fine-tuning processes described above), models competing with a two-step fine-tuning strategy would have to also adapt for the domain difference (i.e. differences in prior and conditional distributions). Second $\mathcal{T}_{inter}$ worth considering is training on Pubmed [20] article-abstracts combinations. While these are not summaries in the stricter sense of the word, this is the closest domain-specific dataset that we could find, and we would like to understand if it adds useful information to a LLM.

4. Models compared

We select LLMs that reduce the amount of overall training required. We select a mix of domain-specific pre-training and general pre-training datasets, and we attempt different $\mathcal{T}_{inter}$s to see how well the resulting models generalize to $\mathcal{T}_T$, namely BioASQ Task 11b Phase B. Hence, the final list of LLMs we are considering are:

- BART - a baseline two-step LLM (without additional fine-tuning) used to establish a baseline for a general domain model without specialized domain knowledge or $\mathcal{T}_{inter}$ fine-tuning
- BioBART - a biomedical two-step LLM (without fine-tuning $\mathcal{T}_{inter}$), used to establish a baseline for an in-domain model
- BART CNN - a baseline LLM three-step LLM with task-specific fine-tuning $\mathcal{T}_{inter}$ but without any deep domain knowledge
- BioBART CNN - a biomedical three-step LLM with task-specific fine-tuning $\mathcal{T}_{inter}$
- BART CNN Pubmed - a general domain three-step LLM fine-tuned for $\mathcal{T}_{inter}$ summarisation task, and then further fine-tuned on a domain-specific $\mathcal{T}_{inter}$ dataset containing Pubmed articles

Based on the data available, we believe that these tasks and LLMs offer the greatest benefit for biomedical summarization, and we limit our selection to 5 models that will participate in the BioASQ competition. We are only considering large models because we want the model to analyze as much context as possible, and therefore having a large model helps to double the context length (1024 tokens vs. 512 tokens). We are using pre-trained BART, BioBART, and BART CNN models available via Huggingface[1]; and we are fine-tuning BART CNN on Pubmed

---
[1]https://huggingface.co/models
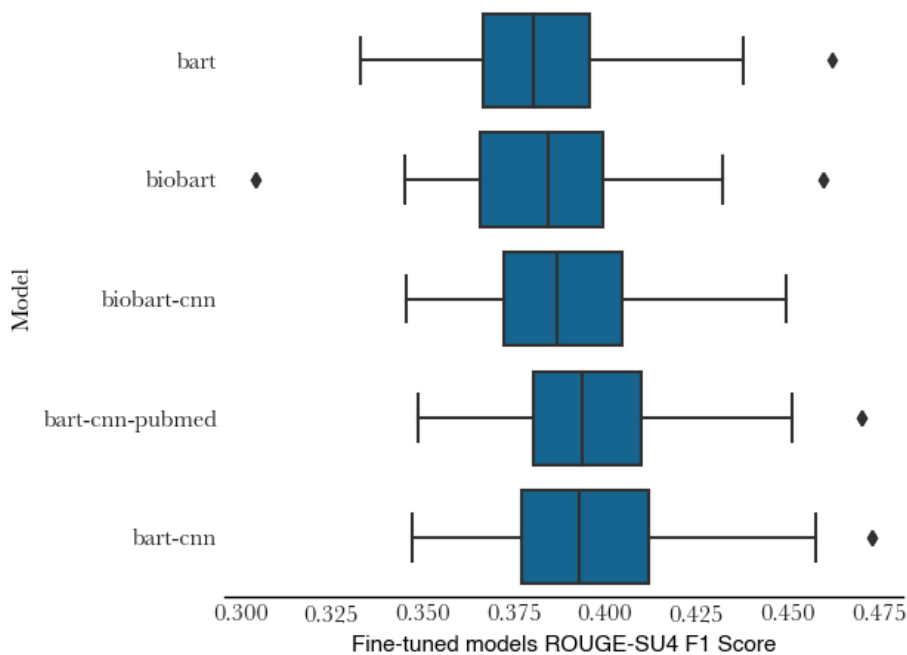
Figure 2: We can see that BioBART outperforms regular BART, which can be attributed to the domain-specific vocabulary; however, the same vocabulary hurts its ability to do task-specific training on the out-of-domain CNN dataset. BART CNN shows a solid performance for a model that was only fine-tuned on a small BioASQ in-domain dataset, and adding an in-domain Pubmed dataset reduces the range of scores a little, making it the best overall model.

data and BioBART on CNN data for one epoch each (our $\mathcal{T}_{inter}$). Subsequently, all models are fine-tuned on the latest complete BioASQ 11 dataset ($\mathcal{T}_T$) for five epochs using a 10-fold cross-validation process. We empirically chose the number of training epochs to maximize the final model scores. Using Pubmed (summarisation) data for fine-tuning BioBART after CNN training didn't offer advantages (Appendix A) and was excluded from the top five models under consideration.

5. Results

Our experiments have revealed a substantial (over 10%) variation in ROUGE [25] score results based on a simple choice of a seed parameter for cross-validation. This indicates that the fine-tuning process is susceptible to changes in data. Future studies should consider which BioASQ information snippets are passed to the model as the input for summarization training. Working with small healthcare question-answering datasets can require a more careful knowledge extraction process [26].

We have experimented with fine-tuning for up to 10 epochs on the $\mathcal{T}_T$, and found that this problem consistently persists across a range of training scenarios. In-domain studies we have

Table 1
ROUGE-SU4 F1 - Mean confidence intervals

model name	mean	CI 95%
BioBART	0.383	(0.373, 0.394)
BART	0.384	(0.376, 0.392)
BioBART CNN	0.39	(0.382, 0.398)
BART CNN	**0.396**	(0.387, 0.406)
BART CNN Pubmed	**0.396**	(0.386, 0.405)

Table 2
ROUGE-SU4 Recall - Mean confidence intervals

model name	mean	CI 95%
BioBART	0.398	(0.386, 0.409)
BART	0.4	(0.392, 0.409)
BioBART CNN	0.415	(0.407, 0.424)
BART CNN	0.42	(0.411, 0.429)
BART CNN Pubmed	**0.422**	(0.41, 0.434)

reviewed show that the generation results can often differ by a minimal margin, significantly lower than the variation in scores we have observed in cross-validation.

To our knowledge, this research is the first to draw attention to this specific problem, and we decided to overcome this by repeating the 10-fold cross-validation training process $\mathcal{T}_T$ four times using a different seed value. Therefore, we effectively report the average of 400 runs for each model (95% t-test confidence interval is given in parentheses), with 100 runs for each seed choice (ten for each fold). We are primarily focused on SU4-F1 scores (Table 1) since they have been shown to correlate with human scores the best [27]. However, ROUGE is recall-oriented; therefore, we also look at Recall results separately (Table 2).

Our experiments (Section 5) suggest that LLMs without domain-specific pre-training show a better capacity for domain-specific text generation. This becomes particularly clear when comparing BART and BioBART results before any additional task-specific fine-tuning, suggesting that BioASQ data is not as similar to Pubmed pre-training data as we would expect based on other results reported on discriminatory tasks. Moreover, we believe that currently a non-domain specific CNN summarization task $\mathcal{T}_{inter}$ is required to accomplish the best results on a BioASQ task. Adding in-domain Pubmed data improves Recall; however, Pubmed data is unsuitable for training for a summarization task from scratch. ROUGE Recall scores (Section 5) show one notable difference, BART CNN has a higher recall, whereas BART CNN Pubmed has a higher precision, likely because the Pubmed training after the task-specific training introduces a task-specific vocabulary to the model.

Overall, LLMs have established some remarkable results in various practical applications. However, since LLMs require task-specific datasets to train to generate text, and such domain-specific datasets are scarce, we need to find ways to overcome these challenges. We have presented an approach that focuses on applications of transfer learning to a domain with limited

task-specific training data.

6. Conclusion and Future Work

In this work, we have observed that task-specific data is critical for generating text in a biomedical domain. Based on our experiments, models without in-domain pre-training are better at summarizing BioASQ data. Unfortunately, our models have achieved fairly modest automated ROUGE scores during BioASQ 11 runs, and we are waiting for the final results to determine how the models have performed overall. The generation process is non-deterministic, and while the answers generated by the models appear sensible, we need better ways to evaluate the candidates.

We have discussed how transfer learning can overcome challenges with data availability. We see a lot of exciting possibilities for using generator models (more specifically paraphrasing, simplification, and rewriting models [28]) for creating synthetic training data, as well as for providing a differentiable loss function which allows sampling a wider space of possible answers without over-penalizing exploration. Abstractive summarization models are trained to generate specific gold sequences, even when they start making errors in the first steps (a problem known as exposure bias [29]). One recent improvement over BART proposes generating multiple candidates and comparing them, showing a new SOTA on several popular summarization datasets [30]. This could address a common shortcoming of autoregressive models, leading to further performance improvements. Another possibility that shows a significant promise would be generating synthetic data to augment BioASQ. This approach has recently shown good results in machine translation [31], and we believe it can be used for other text-generation problems.

References

[1] H. M. Krumholz, S. F. Terry, J. Waldstreicher, Data acquisition, curation, and use for a continuously learning health system, 2016. doi:`10.1001/jama.2016.12537`.

[2] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu,

S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, P. Liang, On the opportunities and risks of foundation models (2021).

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, I. Polosukhin, G. Brain, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, I. Polosukhin, L. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems 2017-Decem (2017). URL: http://arxiv.org/abs/1706.03762.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL: http://arxiv.org/abs/1810.04805. doi:N19-1423.

[5] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettle-moyer, Bart: Denoising sequence-to-sequence pre-training for natural language genera-tion, translation, and comprehension, Association for Computational Linguistics, 2020, pp. 7871–7880. URL: http://arxiv.org/abs/1910.13461https://www.aclweb.org/anthology/2020.acl-main.703. doi:10.18653/v1/2020.acl-main.703.

[6] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, 2013, pp. 1–12. URL: http://arxiv.org/abs/1301.3781.

[7] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, volume 1, 2018. URL: http://arxiv.org/abs/1802.05365. doi:10.18653/v1/n18-1202.

[8] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. B. A. McDermott, Publicly available clinical bert embeddings (2019). URL: http://arxiv.org/abs/1904.03323.

[9] J. Lee, W. Yoon, S. S. Kim, D. Kim, S. S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics (2020) 36 (2019). URL: http://arxiv.org/abs/1901.08746. doi:10.1093/bioinformatics/btz682.

[10] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing (2020). URL: http://arxiv.org/abs/2007.15779.

[11] S. Alrowili, K. Vijay-Shanker, Biom-transformers: Building large biomedical language models with bert, albert and electra, 2021. doi:10.18653/v1/2021.bionlp-1.24.

[12] S. Raza, D. J. Reji, F. Shajan, S. R. Bashir, Large-scale application of named entity recognition to biomedicine and epidemiology, PLOS Digital Health 1 (2022). doi:10.1371/journal.pdig.0000152.

[13] Y. Liu, M. Lapata, Text summarization with pretrained encoders, 2019. URL: http://arxiv.org/abs/1908.08345. doi:10.18653/v1/d19-1387.

[14] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer (2019). URL: http://arxiv.org/abs/1910.10683.

[15] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, S. Yu, Biobart: Pretraining and evaluation of a biomedical generative language model, 2022. doi:10.18653/v1/2022.bionlp-1.9.

[16] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, T. Y. Liu, Biogpt: generative pre-trained transformer for biomedical text generation and mining, Briefings in bioinformatics 23 (2022). doi:10.1093/bib/bbac409.

[17] D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, D. Song, Pretrained transformers

improve out-of-distribution robustness, 2020. doi:`10.18653/v1/2020.acl-main.244`.

[18] A. Radford, K. Narasimhan, I. Sutskever, Gpt: Improving language understanding by generative pre-training, OpenAI (2018).

[19] R. Alec, W. Jeffrey, C. Rewon, L. David, A. Dario, S. Ilya, Language models are unsupervised multitask learners, OpenAI Blog 1 (2019). URL: https://github.com/codelucas/newspaper.

[20] PubMed, About - pubmed, 2021. URL: https://pubmed.ncbi.nlm.nih.gov/about/.

[21] S. J. Pan, Q. Yang, A survey on transfer learning, IEEE Transactions on Knowledge and Data Engineering 22 (2010) 1345–1359. URL: http://ieeexplore.ieee.org/document/5288526/. doi:`10.1109/TKDE.2009.191`.

[22] S. Ruder, Neural transfer learning for natural language processing, 2019.

[23] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, 2016.

[24] K. M. Hermann, T. Kočiský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, volume 2015-January, 2015.

[25] C. Y. Lin, Rouge: A package for automatic evaluation of summaries, 2004.

[26] Y. Liu, S. Chowdhury, C. Zhang, C. Caragea, P. S. Yu, Interpretable multi-step reasoning with knowledge extraction on complex healthcare question answering (2020).

[27] D. Molla-Aliod, C. Jones, Classification betters regression in query-based multi-document summarisation techniques for question answering: Macquarie university at bioasq7b, 2020. URL: http://arxiv.org/abs/1909.00542.

[28] Q. Xu, C. Xu, L. Qu, Alter: Auxiliary text rewriting tool for natural language generation, 2019. doi:`10.18653/v1/d19-3003`.

[29] S. Bengio, O. Vinyals, N. Jaitly, N. Shazeer, Scheduled sampling for sequence prediction with recurrent neural networks, volume 2015-January, 2015.

[30] Y. Liu, P. Liu, D. Radev, G. Neubig, Brio: Bringing order to abstractive summarization, volume 1, 2022. doi:`10.18653/v1/2022.acl-long.207`.

[31] E. Briakou, M. Carpuat, Can synthetic translations improve bitext quality?, volume 1, 2022. doi:`10.18653/v1/2022.acl-long.326`.

A. Additional results
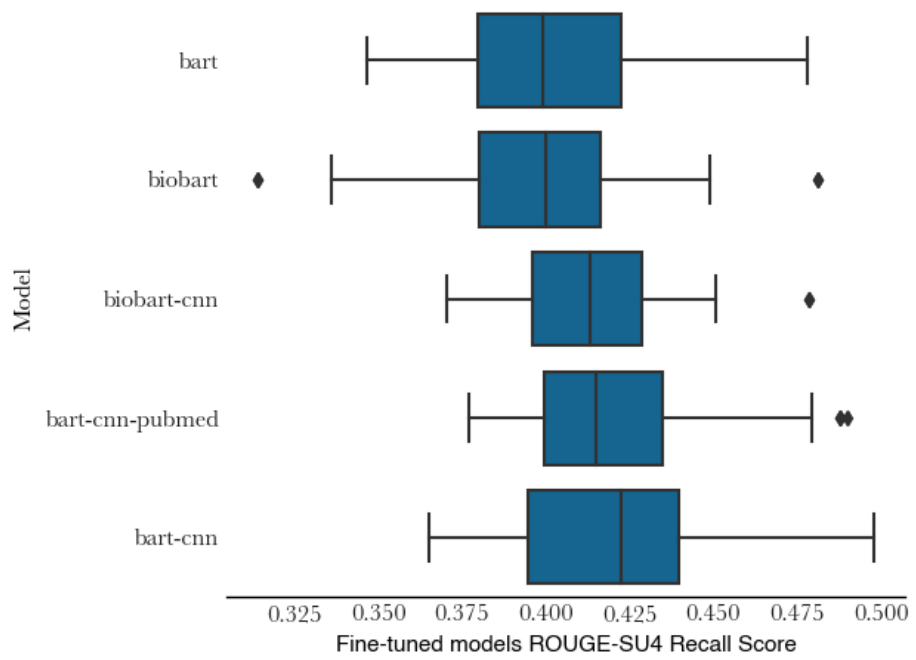
Figure 3: ROUGE SU4-Recall

Figure 4: Biobart CNN + Pubmed: ROUGE4-F1

rouge4

_name_or_path: biobart-cnn-pubmed/

biobart-large${x}: ${mean} σ ${stddev} (${min}, ${max})

_name_or_path: facebook/bart-large

_name_or_path: theojolliffe/bart-large-cnn-finetuned-pubmed

_name_or_path: facebook/bart-large-cnn

_name_or_path: biobart-cnn/

0.00   0.05   0.10   0.15   0.20   0.25   0.30