MDC at BioLaySumm Task 1: Evaluating GPT Models for Biomedical Lay Summarization

Oisín Turbitt and **Robert Bevan** and **Mouhamad Aboshokor**
Medicines Discovery Catapult
{oisin.turbitt, robert.bevan}@md.catapult.org.uk
mouhamadaboshokor@gmail.com

Abstract

This paper presents our approach to the BioLaySumm Task 1 shared task, held at the BioNLP 2023 Workshop. The effective communication of scientific knowledge to the general public is often limited by the technical language used in research, making it difficult for non-experts to comprehend. To address this issue, lay summaries can be used to explain research findings to non-experts in an accessible form. We conduct an evaluation of autoregressive language models, both general and specialized for the biomedical domain, to generate lay summaries from biomedical research article abstracts. Our findings demonstrate that a GPT-3.5 model combined with a straightforward few-shot prompt produces lay summaries that achieve significantly higher relevance and factuality compared to those generated by a fine-tuned BioGPT model. However, the summaries generated by the BioGPT model exhibit better readability. Notably, our submission for the shared task achieved 1st place in the competition.

1 Introduction

Effective communication of scientific ideas is essential for sharing research findings with the general public. While scientific publications serve as an important tool for scientists to share their work, they are primarily intended for other researchers. The use of technical language and the required background knowledge makes these articles difficult for the general public to comprehend, limiting accessibility and research impact (Kuehne and Olden, 2015). The goal of lay summarization is to generate a concise summary of technical texts without using technical language so it can be understood by non-experts. Although certain publishers require authors to provide a lay summary alongside their manuscripts, not all publishers have this requirement, and historical publications are unlikely to include such summaries. Therefore, the development of effective automatic lay summarization becomes crucial in bridging this communication gap (Guo et al., 2021).

GPT-style autoregressive language models have demonstrated impressive ability for natural language generation (NLG) tasks including summarization (Radford et al., 2018; Radford et al., 2019). These models are pre-trained on large corpora and can leverage knowledge gained from this data to generate text. Recent research has rapidly advanced the performance of GPT-style models by substantially scaling the size of these models and their training corpora (Brown et al., 2020), along with instruction tuning and reinforcement learning with human feedback (Ouyang et al., 2022).

Researchers have evaluated GPT models on biomedical-specific natural language understanding (NLU) tasks and found that they underperform due to a lack of domain knowledge (Jimenez Gutierrez et al., 2022; Moradi et al., 2022). Recent research by Luo et al. (2022a) aimed to alleviate this shortcoming by proposing BioGPT, a model based upon GPT-2 (Radford et al., 2019) but specifically trained on biomedical text from PubMed[1]. The authors demonstrate that BioGPT can significantly outperform GPT-2 on biomedical NLU tasks; however, there is currently no evaluation on biomedical natural language generation (NLG) tasks. Therefore, the objective of this work is to evaluate the ability of general-purpose and domain-specific GPT models to generate lay summaries of biomedical research articles as part of the BioLaySumm challenge (Goldsack et al., 2023).

This approach raises an issue. The attention mechanism used in many GPT-style models has quadratic memory scaling with the length of a sequence, which limits the maximum length of input text that can be processed. At the time of writing, the average length of a research article exceeded this limit. Journal articles already contain author-

[1] https://pubmed.ncbi.nlm.nih.gov/

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 611–619

written abstracts that provide a technical summary of the paper and previous work by Goldsack et al. (2022) demonstrated the importance of the abstract in lay summarization. Therefore, we focused on using only the article abstract as the input and leveraging the domain knowledge of these models to generate lay summaries.

We present results for this method for zero-shot and fine-tuned BioGPT and GPT-2 models. Additional experiments are conducted using GPT-3.5 variants accessed via the OpenAI API and we compare their performance with the smaller - but domain-specific - BioGPT model. Our findings indicate that a GPT-3.5 model, combined with a straightforward few-shot prompt, achieves the best performance among the evaluated models. This approach secured 1st place in the shared task.

2 Background

2.1 Autoregressive language models

Autoregressive language models are trained using a conditional generation task, where the model is trained to predict the next token in a sequence given all previous tokens in the sequence. The model is given a sequence of tokens x of length T from a training corpus of N sequences, where each sample consists of a sequence of numeric tokens that map to a natural language vocabulary. The model's training objective is then formulated as maximizing the log-likelihood across the training corpus, given the model's parameters θ:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \sum_{t=1}^{T_i} \log P_\theta(x_{i,t}|x_{i,1:t-1}) \quad (1)$$

The parameters θ are learned through backpropagation and stochastic gradient descent. Once this first phase of training is completed, often referred to as pre-training, the model can be fine-tuned for a specific task using additional data. To achieve this, the input and expected output for the model needs to be converted into a format that is compatible with the training objective in Equation 1. This requires framing the task in natural language and formatting the training data to match this structure.

A popular approach (Brown et al., 2020), and the method used in Luo et al. (2022a), is to train the model on sequences consisting of the information required for the task, called the **source**, along with a natural language description of the task, called the **prompt**, and the expected output, called the **target**. For summarization tasks, this task can then be structured as the text to be summarized as the source, a description of the summarization required as the prompt and an example summary as the target. At inference time, only the source and prompt are provided as input and the model generates the missing summary.

	Sample count		
	Train	Val	Test
PLOS	24,773	1,376	142
eLife	4,346	241	142

	Summary token count		
	Min	Avg	Max
PLOS	203	429±74	773
eLife	10	233±49	599

Table 1: PLOS and eLife dataset statistics.

2.2 Zero-shot and few-shot prompting

The simplest method for querying a language model - known as zero-shot prompting - involves passing a single prompt alongside the source and requires the model to immediately generate the target. Few-shot prompting is a form of in-context learning (Dong et al., 2023) that builds upon zero-shot prompting by providing several source texts and target examples along with a prompt instruction. For lay summarization, these examples would be abstract and lay summary pairs. See Appendix A for example zero and few-shot prompts. While zero and few-shot prompting can produce impressive results (Brown et al., 2020), fine-tuning a model to perform a task is usually a more powerful approach but requires significantly more resources and access to the model parameters.

3 Experiments

3.1 Dataset

The dataset for the shared task is outlined in Goldsack et al. (2022) and Luo et al. (2022b). It consists of two sets of biomedical articles from the eLife and Public Library of Science (PLOS) journals. Table 1 lists the dataset statistics.

For each article, the entire text was provided along with an expert-created lay summary of the article. The articles were split based on the sections contained within and the abstract was extracted.

3.2 Metrics

The generated summaries were evaluated according to three criteria: relevance, readability, and factuality. Relevance was measured by the ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004; Lin and Hovy, 2003; Lin and Och, 2004) and BERTScore (Zhang et al., 2020) metrics. Readability was measured using the Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975) and Dale-Chall Readability Score (DCRS) (Chall and Dale, 1995) metrics. The BARTScore metric (Yuan et al., 2021) - computed using a model fine-tuned for the shared task by the organisers - was employed to measure factuality.

3.3 Experimenting with GPT models

We fine-tuned and evaluated the open-source GPT models using the PyTorch v2.0.1 (Paszke et al., 2019) and Huggingface Transformers v4.28.1 (Wolf et al., 2020) libraries. The following pre-trained models were experimented with: GPT-2_{MEDIUM}[2], GPT-2_{XL}[3], BioGPT[4] and $\text{BioGPT}_{\text{LARGE}}$[5]

To create the training samples used to fine-tune the models, the abstract was used as the source input, a prompt of "Explanation: " was used and the provided lay summary of the article was used as the target. The eLife and PLOS training datasets were combined to form a single training dataset. During inference, the model was given an abstract as the source along with the same prompt used in training. The text generated by the model after the prompt was extracted and used to evaluate the model. Contrary to the fine-tuning work carried out in Luo et al. (2022a), we focused solely on using manually designed prompts instead of the prefix-tuned soft prompts (Li and Liang, 2021) the authors used.

It was decided that due to the resource and energy requirements in training large language models, no hyperparameter optimisation was to be performed for any of the models being evaluated and the same hyperparameters were to be used for all models. Fine-tuning training for all models was performed across 4 Nvidia V100-SXM2 32GB GPUs with PyTorch DDP (Li et al., 2020), using gradient checkpointing and FP16 precision. Each model was fine-tuned for a maximum of 4 epochs with a total batch size of 32. The Adafactor optimiser (Shazeer and Stern, 2018) was used with a peak learning rate of 2e-5. During training, a weight decay of 0.01 was applied and the gradient normal was clipped to a maximum value of 1. The learning rate followed a cosine-decay schedule, with 10% of the total training steps used as warm-up steps. Validation metrics were calculated after every 50 steps, with training stopped early if the validation loss did not decrease after 3 consecutive evaluation steps.

For text generation, contrastive search sampling (Su et al., 2022; Su and Collier, 2023) was used, with hyperparmeters $k = 6$ and $\alpha = 0.5$ for a maximum of 768 new tokens. Again, no hyperparameter selection was carried out for these values and they were set based on manual evaluation of selected generations from the validation set. During inference, all text generation was carried out on a single V100-SXM2 32GB GPU using FP16 precision.

3.4 Experimenting with the OpenAI API

GPT-3.5 exhibits strong few-shot performance across a range of natural language processing tasks (Ouyang et al., 2022) and the OpenAI API enables users to easily query their models for a fee. Given the model's strong performance in NLG tasks and the ease of using the API, we experimented with two GPT-3.5 models[6] for lay summarization: **gpt-3.5-turbo** (the model used by ChatGPT[7], hereafter referred to as the chat model) and **text-davinci-003** (the most powerful instruction-tuned model, hereafter referred to as the instruct model). OpenAI's documentation suggests the chat model offers comparable performance to the instruct model at a fraction of the cost. For this reason we performed the majority of our experiments using the chat model.

We experimented with two different prompting schemes: a simple zero-shot prompt, and few-shot abstract-lay summary pair prompts. When generating abstract-lay summary pair prompts we included as many as would fit in the API call's token limit when combined with the abstract to be summarized and the generated summary (4096 and 4097 tokens for the chat and instruct models, respectively). The length of the generated summary was limited to the maximum summary length observed in the relevant training data set (Table 1). This resulted in eLife and PLOS few-shot prompts containing ∼3-5 and

[2]https://huggingface.co/gpt2-medium
[3]https://huggingface.co/gpt2-xl
[4]https://huggingface.co/microsoft/biogpt
[5]https://huggingface.co/microsoft/BioGPT-Large

[6]https://platform.openai.com/docs/models/gpt-3-5
[7]https://openai.com/blog/chatgpt

∼3-7 example pairs respectively.

The temperature parameter - which controls the entropy of the token probability distribution during text generation - was tuned using the eLife dataset only to reduce costs. The performance was consistent for each temperature setting; we selected a temperature of 0.3 for the final evaluation.

4 Results and Discussion

Table 2 lists the combined eLife and PLOS validation set performance metrics, with the few-shot instruct model performing best overall. As expected, fine-tuning the BioGPT and GPT-2 models greatly enhances the relevance, readability, and factuality of the generated summaries. Both fine-tuned BioGPT models outperform the GPT-2 models across nearly all metrics, demonstrating the beneficial impact of leveraging biomedical domain knowledge gained during pre-training for text generation. The smaller BioGPT and GPT-2_{MEDIUM} models exhibit better readability metrics than the large model variants, however the relevance and factuality metrics are significantly worse. Interestingly, the zero-shot performance of both BioGPT models is considerably inferior to that of the zero-shot GPT-2 models. This discrepancy may be attributed to the smaller dataset used for pre-training the BioGPT models which could reduce the zero-shot generalisation performance of the model, but this would need further investigation.

Few shot prompting of the instruct model significantly improved its performance and this model achieved the best results for the factuality and relevance metrics. The remarkable increase in performance of the few-shot versus the equivalent zero-shot models demonstrates the large language models' ability for in-context learning. Interestingly, the performance of the fine-tuned BioGPT$_{\text{LARGE}}$ model is not far off the text-davinci-003 despite being 1/100$^{\text{th}}$ the size of the GPT-3.5 model and even outperformed chat and instruct models in their zero-shot setting. This shows the benefit of domain knowledge for domain-specific natural language generation tasks but also alludes to potential performance gains of scaling up the size of domain-specific language models.

The chat model performed significantly better than the instruct model when using the zero-shot prompt. This may be because the zero-shot prompt is in a conversational style, which the chat model was trained for, or additional training to the chat model that the instruct model did not receive. We are limited to conjecture as the training details and differences of the chat model from the instruct model are largely unknown as they have not been published by OpenAI.

Due to the superior performance on the validation set scores, we chose the instruct model for the final test evaluation. The test set scores are listed in Table 3 alongside the fine-tuned BioGPT$_{\text{LARGE}}$ test scores and the test scores of the BART model proposed by the shared task organisers (Goldsack et al., 2022). The few-shot instruct model outperforms both of these models, achieving either the best or second best score for nearly all metrics. However, the fine-tuned BioGPT$_{\text{LARGE}}$ model did generate summaries with the best readability at the cost of worse factuality. We submitted the few-shot instruct model test set results to the shared task competition and out of 21 participants, the model came in 1$^{\text{st}}$ place overall. Across the evaluation criteria, the model came in 3$^{\text{rd}}$ place for relevance and factuality but only achieved 10$^{\text{th}}$ place on readability.

It would be valuable to compare the performance of the proposed approaches with other recently open-sourced large language models such as the biomedical-domain BioMedLM[8], the scientific-domain Galactica (Taylor et al., 2022) or the general-domain LLaMA (Touvron et al., 2023). Another interesting future research avenue is available with the longer context models. The emergence of models such as GPT-4 (OpenAI, 2023), PALM-2 (Anil et al., 2023) and Claude[9] offers significantly increased context lengths compared to the models assessed in this paper. Using these models, it would be possible to input more sections such as the introduction from a paper into the few-shot prompts, potentially leading to substantial enhancements in performance as these sections provide additional background information. Furthermore, the performance of the BioGPT$_{\text{LARGE}}$ model has room for improvement. All hyperparameters were fixed during training of the model due to resource constraints so tuning of the training and generation hyperparameters may benefit model performance. The hyperparameters for all models were also not tuned for a specific dataset and separately tuning these for each dataset may improve summary quality. Advanced prompting techniques have been

[8] https://huggingface.co/stanford-crfm/BioMedLM

[9] https://www.anthropic.com/index/introducing-claude

Model	Tuning	# Params	RG-1↑	RG-2↑	RG-L↑	BERTScore↑	FKGL↓	DCRS↓	BARTScore↑
GPT-2_{M}	0S	355M	0.289	0.060	0.269	0.820	13.307	9.245	-3.836
GPT-2_{M}	FT	355M	0.370	0.084	0.348	0.830	12.725	**8.745**	-3.523
BioGPT	0S	347M	0.070	0.017	0.061	0.683	**8.622**	11.556	-6.063
BioGPT	FT	347M	0.416	0.115	0.388	0.846	11.885	9.614	-3.047
GPT-2_{XL}	0S	1.6B	0.295	0.066	0.275	0.816	12.847	9.435	-3.913
GPT-2_{XL}	FT	1.6B	0.360	0.085	0.339	0.832	12.684	8.818	-3.638
BioGPT$_{LARGE}$	0S	1.5B	0.274	0.075	0.252	0.811	13.597	11.422	-3.985
BioGPT$_{LARGE}$	FT	1.5B	0.441	0.134	0.411	0.851	12.139	9.808	-2.832
gpt-3.5-turbo	0S	Unknown	0.391	0.112	0.356	0.853	13.196	10.460	-3.050
gpt-3.5-turbo	FS	Unknown	0.418	0.127	0.381	0.856	13.614	10.746	-2.839
text-davinci-003	0S	175B	0.346	0.098	0.312	0.848	13.502	10.708	-3.314
text-davinci-003	FS	175B	**0.460**	**0.144**	**0.424**	**0.861**	13.514	10.402	**-2.029**

Table 2: The performance of the evaluated models on the combined PLOS and eLife validation sets. The best score for each metric is highlighted in bold and the second best score is underlined. 0S is short for zero-shot, FS is short for few-shot and FT is short for fine-tuned.

Model	RG-1↑	RG-2↑	RG-L↑	BERTScore↑	FKGL↓	DCRS↓	BARTScore↑
BART (Baseline)	0.470	0.145	0.437	0.864	**12.069**	10.249	**-0.831**
BioGPT$_{LARGE}$	0.434	0.130	0.403	0.851	12.681	**10.036**	-2.876
text-davinci-003	**0.482**	**0.155**	**0.449**	**0.871**	12.937	10.206	-1.177

Table 3: The test set performance of the fine-tuned BioGPT$_{LARGE}$ model, the few-shot text-davinci-003 model and a BART model proposed by the task organisers (Goldsack et al., 2022).

shown to improve performance over manually designed prompts and may enhance performance (Liu et al., 2022; Li and Liang, 2021).

While the capability shown in these results is intriguing, this method is not yet mature enough to be relied upon to generate lay summaries. GPT models are known to hallucinate factual information (Ji et al., 2023) and detecting these hallucinations in text is difficult for both machines and the untrained eye. If the general public were to depend on these generated lay summaries for comprehending scientific research, the confident dissemination of erroneous information would be actively detrimental. Further work is required to verify the factuality of the generated lay summaries using human experts before this approach could be safely used.

5 Conclusion

This paper presents an evaluation of general-purpose and biomedical domain-specific autoregressive language models for generating lay summaries from the abstracts of scientific articles for the BioLaySumm shared task at BioNLP 2023. Our findings demonstrate that the biomedical domain-specific model, BioGPT, outperforms general-purpose GPT-2 models when fine-tuned for lay summary generation. We also explored the effectiveness of zero and few-shot prompting for generating lay summaries using OpenAI's GPT-3.5 models. While the zero-shot performance of these models is worse than the fine-tuned BioGPT$_{LARGE}$ model, we discovered that using the text-davinci-003 coupled with a few-shot prompt yielded the best results among all the language models tested. This approach was selected as our submission for the shared task, achieving the overall 1st place submission and 3rd place in both relevance and factuality metrics. To further enhance the performance of this approach, we anticipate that advanced prompting methods, evaluation of additional models, and utilization of models with longer context lengths could be beneficial. Despite the promising results obtained in this study, it is essential to conduct further research to validate the factuality of the generated lay summaries using human experts before practical application.

Limitations

Our best results were obtained using a few-shot prompt of the text-davinci-003 model from OpenAI. While the technical barriers to this method are very low due to the ease of use of the model API, the cost of querying the API can become prohibitively expensive and this limited our own experiments with the model[10]. This cost would rise significantly more if the text-davinci-003 was fine-tuned to perform these lay summaries. If this is a concern, then using the open-sourced BioGPT models may be beneficial. It should be noted that performing the fine-tuning process is itself expensive and requires access to high-end GPUs. Practitioners should investigate parameter-efficient fine-tuning techniques (Hu et al., 2021) if access to these GPUs is an issue.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

J.S. Chall and E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated lay language summarization of biomedical scientific reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 160–168.

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.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical IE? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval

[10]Generating validation results for both GPT-3.5 cost $115 on the 1st June 2023

Technical Training Command Millington TN Research Branch.

Lauren M. Kuehne and Julian D. Olden. 2015. Lay summaries needed to enhance science communication. *Proceedings of the National Academy of Sciences*, 112(12):3585–3586.

Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, and Soumith Chintala. 2020. Pytorch distributed: Experiences on accelerating data parallel training. *Proc. VLDB Endow.*, 13(12):3005–3018.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

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.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pages 150–157.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022a. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2022b. Readability controllable biomedical document summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4667–4680, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Milad Moradi, Kathrin Blagec, Florian Haberl, and Matthias Samwald. 2022. Gpt-3 models are poor few-shot learners in the biomedical domain.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Yixuan Su and Nigel Collier. 2023. Contrastive search is what you need for neural text generation.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. *arXiv preprint arXiv:2202.06417*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

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.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *CoRR*, abs/2106.11520.

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

A Example prompts

Zero-shot prompt
"Provide a lay summary of the following research abstract:"

Few-shot prompt
"**Abstract:** "The thiamine pyrophosphate (TPP) riboswitch is a cis-regulatory element in mRNA that modifies gene expression..." **Lay summary:** "When a gene is switched on , its DNA is first copied to make a molecule of messenger ribonucleic acid (mRNA)..." **Abstract:** "In Heliconius butterflies , wing colour pattern diversity and scale types are controlled by a few genes of large effect.." **Lay summary:** "Heliconius butterflies have bright patterns on their wings that tell potential predators that they are toxic..." **Abstract:** "Diverse interactions among species within bacterial colonies lead to intricate spatiotemporal dynamics, which can affect..." **Lay summary:** "Communities of bacteria and other microbes live in every ecosystem on Earth, including in soil..." **Abstract:** "The ability to recognize foreign double-stranded (ds) DNA of pathogenic origin in the intracellular environment is an essential..." **Lay summary:**"

Table 1: Example zero-shot and few-shot prompts used with the GPT-3.5 models.