DeakinNLP at BioLaySumm: Evaluating Fine-tuning Longformer and GPT-4 Prompting for Biomedical Lay Summarization

Huy Quoc To, Ming Liu, Guangyan Huang
School of Information Technolgy, Deakin University, Australia
{q.to, m.liu, guangyan.huang}@deakin.edu.au

Abstract

This paper presents our approaches for the BioLaySumm 2024 Shared Task. We evaluate two methods for generating lay summaries based on biomedical articles: (1) fine-tuning the Longformer-Encoder-Decoder (LED) model, and (2) zero-shot and few-shot prompting on GPT-4. In the fine-tuning approach, we individually fine-tune the LED model using two datasets: PLOS and eLife. This process is conducted under two different settings: one utilizing 50% of the training dataset, and the other utilizing the entire 100% of the training dataset. We compare the results of both methods with GPT-4 in zero-shot and few-shot prompting. The experiment results demonstrate that fine-tuning with 100% of the training data achieves better performance than prompting with GPT-4. However, under data scarcity circumstances, prompting GPT-4 seems to be a better solution.

1 Introduction

The task of summarization has witnessed the development based on pre-trained language models. More recently, the superiority of large language models (LLMs) has been demonstrated on a wide range of natural language processing (NLP) tasks (Minaee et al., 2024; Zhao et al., 2023). In the BioLaySumm 2024 shared task (Goldsack et al., 2024), the competition focuses on generating summaries for biomedical research articles that are easily understandable by the general public. These summaries are usually known as "lay summaries".

Recently, the study of the summarization task using generative models has increased for both general domains (Koh et al., 2022b; Zhao et al., 2020) and biomedical text (Liu et al., 2023a). Additionally, according to Goldsack et al. (2022), each article generally has more than 10,000 words. Many pre-trained language models have been developed to handle such long text (Koh et al., 2022a). In this paper, we implement the Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020) as an approach for Biolaysumm shared task, as its performance has been demonstrated in (Liu et al., 2023b; Wu et al., 2023).

In this paper, we present a comparison between the performance of the fine-tuned LED model on 50% and 100% of the training set. Additionally, we evaluate GPT-4 (OpenAI et al., 2024) on zero-shot and few-shot prompting for this Shared Task. Our aim is to investigate how a fine-tuned model and a large language model such as GPT-4 perform in lay summarization biomedical text. This study focuses on three aspects: performance, training time, and computational cost. Our contributions are as follows.

- We fine-tune LED model on different amount of data to evaluate how it affects the performance of the LED model in biomedical lay summarization task.
- Secondly, we evaluate GPT-4 on zero-shot and few-shot prompting to investigate how the in-context learning capability of this model. Our results show that, in the eLife dataset, the GPT-4 few-shot prompting method outperforms the fine-tuned LED model.

In the following sections, we briefly analyze the datasets, describe our methods in detail, showcase the experiment settings, and present our results, findings, and conclusion.

2 Datasets

The task is evaluated on two datasets: PLOS and eLife (Goldsack et al., 2022). Both datasets contain biomedical articles and a lay summary manually written for each article. To first understand the evaluation datasets, we proceed tokenizing the input and the output text on two datasets using tokenizer from LED model (Beltagy et al., 2020). We sum-

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 748–754
August 16, 2024. ©2024 Association for Computational Linguistics

marize the statistics of the PLOS and eLife dataset in Table 1.

Dataset	Article(#Tokens)			Summ.(#Tokens)	
	Train	Val	Test	Train	Val
PLOS	9,851	9,924	9,978	263	279
eLife	12,321	12,753	11,967	435	445

Table 1: The mean number of tokens of input and output text in PLOS and eLife datasets. **Summ.** is the abbreviation for lay summary.

According to (Goldsack et al., 2024) and Table 1, while PLOS has more instances of biomedical papers than the eLife dataset, and the length of both input and output text in eLife is longer than PLOS. We also notice that the maximum number of tokens for input text is 28,561 for PLOS and 34,612 tokens in eLife.

3 Evaluation Metrics

In this shared task, the generated summaries are evaluated on three aspects and ten metrics accordingly:

- **Relevance**: ROUGE-1 (R-1), ROUGE-2 (R-2) and ROUGE-L (R-L) (Lin, 2004) and BERTScore (Zhang et al., 2020).
- **Readability**: Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975) and Dale-Chall Readability Score (DCRS) (Chall and Dale, 1995), Coleman-Liau Index (CLI), and LENS (Maddela et al., 2023).
- **Factuality** - AlignScore (Zha et al., 2023), and SummaC (Laban et al., 2022).

The objective of the evaluation is to maximize the Relevance, Factuality and LENS in Readability scores and minimize FKGL, DCRS, and CLI scores.

4 Preliminary Study

Due to the fact that each instance in both datasets is lengthy and may contain a large amount of irrelevant information to generate lay summaries, we perform a heuristic evaluation on the validation sets. We are aware that each article has at least an abstract and a conclusion paragraph. We evaluated the abstract, the conclusion part and other parts of each article with the lay summaries on the **Relevance** aspect. Table 2 shows that in both cases,

Datasets	Section	R-1	R-2	R-L	BertScore
PLOS	Abs.	**0.502**	**0.199**	**0.466**	**0.871**
	Con.	0.154	0.039	0.146	0.803
	Others	0.084	0.041	0.081	0.832
eLife	Abs.	**0.319**	**0.071**	**0.293**	**0.839**
	Con.	0.162	0.026	0.156	0.782
	Others	0.097	0.033	0.095	0.820

Table 2: Analysis on Relevance aspect of the abstract, conclusion and the rest of the content with the lay summaries.

the abstracts written by the author of each article contain the most similar information. These abstracts are likely to be used as the base knowledge when creating the lay summaries. Additionally, the conclusion parts also achieve competitive scores, which indicates that they have potential to be used as sources to generate lay summaries.

5 Experiments

Based on the results of our preliminary study, we first extract the abstract and conclusion paragraph from the original articles. We then perform the fine-tuning process and prompting GPT-4 using the combination of abstract and conclusion from the original articles.

5.1 Fine-tuning LED model

We fine-tune LED model on each dataset individually using 50% and 100% of the training set. We randomly select 50% of the training instances. Fine-tuning processes are performed on Colab Pro [1] using the L4 GPU (22GB VRam). We employ the base version (41M parameters) of the LED model via Huggingface[2], which can process up to 16,384 tokens. In the experiment, the batch size is set to 2 due to the limitation of the GPU VRam, and we train for 2 epochs and set the learning rate to 1e-5. For the PLOS dataset, we set the maximum token at 10,000 for input, and the maximum output sequence length is 400 tokens. Since the eLife dataset has longer input and output sequence lengths, we set the maximum input token to 14,000 tokens, and the output is 600 tokens. These adjustments are made to accommodate the length of the lay summary in each dataset.

[1] https://colab.research.google.com/

[2] https://huggingface.co/docs/transformers/en/model_doc/longformer

Model	R-1	R-2	R-L	BertScore	FKGL	DCRS	CLI	LENS	AlignScore	SummaC
					PLOS					
LED (50%)	0.472	0.157	0.426	0.864	14.459	11.431	15.781	56.053	0.818	**0.741**
LED (100%)	**0.472**	**0.163**	**0.431**	**0.865**	14.299	11.367	15.520	57.090	**0.819**	0.739
GPT-4 zs	0.420	0.114	0.385	0.857	14.648	10.556	15.456	70.621	0.646	0.485
GPT-4 fs	0.431	0.123	0.402	0.860	**14.210**	**10.530**	**15.380**	**70.781**	0.711	0.589
					eLife					
LED (50%)	0.456	0.121	0.435	0.843	9.456	7.760	10.351	67.392	0.631	0.601
LED (100%)	0.461	0.121	0.441	0.848	**9.448**	**7.752**	**10.345**	68.453	0.653	**0.617**
GPT-4 zs	0.465	0.101	0.431	0.847	15.320	10.707	16.641	68.769	0.656	0.477
GPT-4 fs	**0.493**	**0.121**	**0.457**	**0.851**	14.626	10.145	15.435	**70.732**	**0.672**	0.497

Table 3: The performance of the evaluated models on the PLOS and eLife private test sets. The best score for each metric is highlighted in bold, and the second-best score is underlined. ZS is short for zero-shot and FS is short for few-shot.

Model	R-1	R-2	R-L	BertScore	FKLG	DCRS	CLI	LENS	AlignScore	SummaC
BART (Baseline)	0.470	0.140	0.436	**0.862**	**12.035**	**10.147**	**13.485**	48.096	**0.779**	**0.703**
Final Submission	**0.482**	**0.142**	**0.444**	0.858	14.462	10.755	15.477	**63.912**	0.745	0.618

Table 4: Our final submission is the combination of fine-tuned 100% training set LED model on PLOS dataset and GPT-4 few-shot prompting on eLife dataset.

5.2 Prompting GPT-4

GPT-4 demonstrates strong performance on few-shot settings in multiple NLP tasks (Liu et al., 2023c). In our experiments, we access GPT-4 through OpenAI APIs[3]. To save cost, we choose **gpt-4-turbo-preview** version to generate lay summaries. We evaluated GPT-4 in two settings: zero-shot and few-shot prompting. In zero-shot prompting, we directly pass the extracted input to GPT-4 and generate the lay summaries. When creating prompts in few-shot settings, we randomly pick the source-target pairs from the validation set and use them as examples for GPT-4. Since the maximum tokens that GPT-4 can take are 128,000 token, we incorporate as many as possible within the token constraints of the API calls. As the results, PLOS and eLife few-shot prompts contain 4 and 3 example pairs, respectively. The maximum lay summary length is set to 400 tokens and 600 tokens, respectively, for PLOS and eLife. We present an example of a zero-shot prompt and a few-shot prompt in Appendix A.

6 Results

In this section, we list our results on the private test set. The scores are retrieved through the Codabench page of the shared task and reported in Table 3.

PLOS The results clearly demonstrate that fine-tuning the LED model achieves the best performance on relevance and factual aspects. To our surprise, GPT-4 outperforms LED in readability. The FKGL score of the fine-tuned LED model with 100% train set achieves the second best results. However, for other readability metrics, the performance of LED models is worse than GPT-4 prompting. In particular, the gap in the LENS score is noticeably high. The gap is around 13.6 percentage points when comparing the fine-tuned version of LED (100%) with GPT-4 few-shot prompting. Meanwhile, compared to the results of the GPT-4 few-shot prompting, the fine-tuned LED model with full training data outperforms by 0.041, 0.039, 0.029, 0.005, 0.108, and 0.150 on R-1, R-2, R-L, BERTScore, AlignScore, and SummaC, respectively. It seems that the improvement of the best fine-tuned LED on those scores can be considered marginal.

eLife On the eLife dataset, it is surprising that GPT-4 outperforms fine-tuned LED model in generating more accurate summaries. However, the difference in readability is significant, as GPT-4 achieves lower scores on FKGL, DCRS, and CLI compared to LED models. The gaps between GPT-4 and LED model on these three metrics, respec-

[3] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

tively, are 5.178, 2.393, 5.090. Whereas, the differences that GPT-4 few-shot prompting creates compared to LED (100%) fine-tuned version on R-1, R-2, R-L, BertScore, LENS, and AlignScore, respectively, are 0.032, 0, 0.016, 0.003, 2.279, and 0.019. It is no doubt that on eLife dataset, prompting GPT-4 generates better lay summaries in terms of Relevance and Factuality.

Based on the above results, we made our final submission to the shared task by combining the results of the fine-tuned LED model with 100% training data from PLOS and GPT-4 few-shot prompts in the eLife dataset. We compare our submission with the BART baseline (Goldsack et al., 2024) in Table 4. It shows that our results surpass the baseline on the R-1, R-2, R-L, and LENS scores. Remarkably, our LENS score is higher than BART baseline by 15.816%. Although in the other metrics, our results are a bit lower than baseline, we argue that the scores are still competitive and the gap is marginal.

7 Discussion

The results demonstrate that traditional fine-tuning can produce summaries with accurate keywords and context rather than prompting. LED model also creates less hallucination than LLMs, because it achieves better Factuality scores. However, fine-tuning is less effective in making the summaries simpler and easier to understand.

Furthermore, we believe that fine-tuning LED model on eLife is less efficient than on PLOS dataset because of the size of eLife dataset. Furthermore, the text in eLife dataset is also longer than PLOS. Therefore, it is likely that LED model is not able to capture the keywords and learn enough context on eLife. Hence, GPT-4's performance is slightly better in this case.

8 Performance Versus Cost

In this section, we discuss the trade-off between model performance and costs. In our analysis, the costs include training time, computational cost, and prompting cost. We summarize our comparison in Table 5. We first rank the performance of each method based on the results in Table 3. Next, we evaluate four methods based on the number of training hours, the costs of training, inference, and prompting. Since the PLOS dataset has more instances in the training set than eLife, it undoubtedly takes more time and more costly to train LED models on PLOS. In Colab Pro[4], it costs around 5 computational units per hour. Hence, to calculate the total computational cost, we simply multiply 5 by the training time.

Model	**#Rank**	**Training**	**Cost**
	PLOS		
LED (50%)	*2nd*	8 hrs	40 units
LED (100%)	*1st*	20 hrs	100 units
GPT-4 zs	*4th*	0 hr	10$
GPT-4 fs	*3rd*	0 hr	20$
	eLife		
LED (50%)	*4th*	4 hrs	20 units
LED (100%)	*2nd*	8.5 hrs	42.5 units
GPT-4 zs	*3rd*	0 hr	20$
GPT-4 fs	*1st*	0 hr	30$

Table 5: The comparision between four approaches on two datasets. The cost for fine-tuning is referred to computation units and cost for GPT-4 is referred to prompting cost using OpenAI APIs.

On the other hand, we directly prompt GPT-4 without further fine-tuning the model. Therefore, we only report the prompting cost in two data sets. As mentioned in Table 1, the length of each instance in the eLife test set is longer than PLOS, and it costs more to generate the lay summaries. In the few-shot prompting setting, it also costs more because we include more tokens in the queries for example.

Through our result analysis and cost-effective study, it demonstrates that GPT-4 prompting cost us more on querying, however it takes less time then fine-tuning and still achieves competitive results. Especially, in the situation where we have less training data (such as in eLife case), GPT-4 can outperform fine-tuned LED model.

9 Conclusion

This paper details our approach to the BioLay-Summ 2024 shared task, comparing traditional fine-tuning of the Longformer-Encoder-Decoder (LED) model and few-shot prompting with GPT-4 for generating lay summaries of biomedical articles. Our results indicate that the fine-tuned LED excels on the PLOS dataset, while GPT-4's few-shot prompting outperforms LED on the eLife dataset, highlighting GPT-4's advantage in data scarcity scenarios. Future work may explore self-evaluation meth-

[4] In 2024, 100 computational units cost around 15$ on Colab Pro.

ods and cost-reduction strategies for fine-tuning using parameter-efficient techniques.

10 Limitations

Our methodology relies exclusively on OpenAI APIs for generating summaries using GPT-4, which presents minimal technical challenges. However, the costs associated with API requests can quickly escalate to prohibitive levels, limiting our ability to conduct extensive experimental work with the model. Implementing proprietary LLMs such as GPT-4 also has the limitations of reproducing the results. In addition, due to computational cost and time constraints, we were unable to fine-tune the LED model for more epochs, potentially impacting the overall performance.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

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

Tomas Goldsack, Carolina Scarton, Matthew Shardlow, and Chenghua Lin. 2024. Overview of the biolaysumm 2024 shared task on the lay summarization of biomedical research articles. In *The 23rd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Bangkok, Thailand. 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.

Peter Kincaid, Robert P. Fishburne, 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.

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022a. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Comput. Surv.*, 55(8).

Huan Yee Koh, Jiaxin Ju, He Zhang, Ming Liu, and Shirui Pan. 2022b. How far are we from robust long abstractive summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2682–2698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

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.

Ming Liu, Dan Zhang, Weicong Tan, and He Zhang. 2023a. DeakinNLP at ProbSum 2023: Clinical progress note summarization with rules and language ModelsClinical progress note summarization with rules and language models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 491–496, Toronto, Canada. Association for Computational Linguistics.

Quancheng Liu, Xiheng Ren, and V.G.Vinod Vydiswaran. 2023b. LHS712EE at BioLaySumm 2023: Using BART and LED to summarize biomedical research articles. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 620–624, Toronto, Canada. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023c. 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.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. Lens: A learnable evaluation metric for text simplification. *Preprint*, arXiv:2212.09739.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *Preprint*, arXiv:2402.06196.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Yu-Hsuan Wu, Ying-Jia Lin, and Hung-Yu Kao. 2023. IKM_Lab at BioLaySumm task 1: Longformer-based prompt tuning for biomedical lay summary generation. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 602–610, Toronto, Canada. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

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

Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang, and Gholamreza Haffari. 2020. Summpip: Unsupervised multi-document summarization with sentence graph compression. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1949–1952, New York, NY, USA. Association for Computing Machinery.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

A Example prompts on GPT-4

Zero-shot prompt	
role	*system*
content	"Write a lay summary using the following research abstract and conclusion."
role	*user*
content	"Lung-resident (LR) mesenchymal stem and stromal cells (MSCs) are key elements of the alveolar niche and fundamental regulators of homeostasis and regeneration..."
Few-shot prompt	
role	*system*
content	"Write a lay summary using the following research abstract and conclusion."
role	*user*
content	"Gene expression varies widely between individuals of a population , and regulatory change can underlie phenotypes of evolutionary and biomedical relevance..."
role	*assistant*
content	"Messenger RNAs carry the instructions necessary to synthesize proteins that do work for the cell..."
role	*user*
content	"The live attenuated simian immunodeficiency virus (LASIV) vaccine SIVnef is one of the most effective vaccines..."
role	*assisstant*
content	"Annually, more than two million people are infected with HIV , the virus that causes AIDS..."
role	*user*
content	"Mucosal infections with Candida albicans belong to the most frequent forms of fungal diseases..."
role	*assisstant*
content	"The opportunistic pathogen Candida albicans is a major risk factor for immunosuppressed individuals..."
role	*user*
content	"Lung-resident (LR) mesenchymal stem and stromal cells (MSCs) are key elements of the alveolar niche and fundamental regulators of homeostasis and regeneration..."

Table 6: Example of zero-shot prompt and few-shot prompt for GPT-4.