CPSC477/577-SP2024 Final Report: Biomedical Lay Summarization

Xincheng Cai
Department of Graduate School of Arts and Sciences
Yale University
New Haven, CT, 06510
xincheng.cai@yale.edu

Mengmeng Du
Department of Graduate School of Arts and Sciences
Yale University
New Haven, CT, 06510
mengmeng.du@yale.edu

Abstract

Biomedical research articles contain vital information for a wide audience, yet their complex language and specialized terminology often hinder comprehension for non-experts. Inspired by the BIONLP 2024 workshop, we propose a NLP solution to generate lay summaries, which are more readable to diverse audiences. We implemented two transformer-based models, specifically BART and BART-PubMed. Our study investigates the performance of these models across different biomedical topics and explores methods to improve summarization quality through definition retrieval from Webster Medical Dictionary. By enhancing the readability of biomedical publications, our work aims to promote knowledge accessibility to scientific information. *Code is available at:* https://github.com/Cyngua/BioLaySum-Project-2024.

1 Introduction

Biomedical publications encompass the latest research on prominent health-related topics, ranging from common illnesses to global pandemics. Consequently, their content often piques the interest of a wide array of audiences, including researchers, medical professionals, journalists, and even members of the general public. However, these articles tend to assume a certain degree of background knowledge and employ domain-specific language, rendering them challenging to comprehend for those lacking the requisite expertise. This situation results in the specific knowledge being confined within the direct community.(1)

To address this issue, abstractive summarization of biomedical articles presents an excellent opportunity. It can generate summaries that are more readable, containing more background information and less technical terminology, thereby becoming more accessible to non-expert audiences. This type of summarization is referred to as a "lay summary."

Preprint. Under review.

Our paper draws inspiration from the shared task of the BIONLP 2024 workshop: SHARED TASK ON THE LAY SUMMARIZATION OF BIOMEDICAL RESEARCH ARTICLES. This task aims to train a model to generate a more readable lay summary, given the article's abstract and main text as input. To achieve this, we are considering the deployment of transformer-based models, including the Bart and Bart-PubMed model, which is the Bart model fine-tuned on the PubMed dataset. We will compare these two models using comprehensive evaluation metrics. Considering that different topics within biomedical publications may exhibit varying performance, we will train our model on each subset of topics. Finally, we will attempt to improve the model's performance through data preprocessing.

Through our project, we aim to enhance the readability of biomedical publications and help broaden the accessibility of technical texts to non-specialist audiences.

2 Related Work

2.1 Pretrained language model for various tasks

Pre-trained language models (PLMs) are language models that are pretrained on large text corpus in self-supervised ways. Finetuning PLMs for downstream tasks has become a common pipeline to solve NLP problems(2). Models like BERT(3) and RoBERTa (4) have emerged as standard choices for text representation. These autoencoding transformers can be finetuned with additional layers for downstream tasks such as question answering and language inference(3). Nevertheless, the encoder-only architecture of these models limits direct text generation. The GPT models, in contrast, mainly deploy decoder-only transformer architecture(5; 6; 7).

Text summarization belongs to Seq2Seq tasks, which involve input sequences being transformed into output sequences. Some state-of-art Seq2Seq models include the Vanilla Seq2Seq(8) and attention-based Seq2Seq (9) and the transformer (encoder-decoder)(10), which are all initially designed for machine translation. Later on, more models based on encoder-decoder structure in transformer emerged, the most important ones among which include T-5(11) and BART(12).

2.2 Biomedical domain-specific language models

The biomedical natural language processing (NLP) community has increasingly focused on utilizing PLMs to extract and analyze biomedical information across different levels of abstraction(13). Notably, BioBERT(14) and SciBERT(15) are variants of the general BERT model, trained specifically on scientific research corpora sourced from databases like PubMed and semantic scientific papers. Other PLMs, such as ClinicalBERT(16) and Clinical XLNet(17), are pretrained on clinical notes derived from clinical databases. Recent developments have introduced specialized PLMs aimed at addressing generative tasks, for instance, BioBART(18), BioGPT(19) and PubMedGPT. Especially for biomedical text summarization, BART based models pretrained on biomedical domain knowledge, like BioBART, have better performance than directly fine-tuning BART(13).

3 Dataset

We utilized the eLife biomedical journals as our dataset, an open-access peer-reviewed journal with a specific focus on biomedical and life sciences. The dataset is provided by the shared task in .jsonl format, where each line represents a JSON object with the fields outlined in Table1. The datasets have been pre-split for model training and validation, consisting of 4,346 instances earmarked for training and 241 for validation. Furthermore, leveraging the keywords within the dataset, we partitioned the training

dataset into 21 subsets. From these subsets, we focused on the first six categories with the highest article count, as illustrated in Table2 below.

Table 1: Dataset format

Column	Description
Lay summary: string	Article lay summary
Article: string	Article main text
Headings: list	Article headings
Keywords: list	Topic of the article
ID: string	Article ID

Table 2: Keyword Subsets

Keyword	Articles
neuroscience	1240
cell biology	922
developmental biology	553
biochemistry and chemical biology	505
structural biology and molecular biophysics	480
microbiology and infectious disease	420

4 Methods

4.1 Technical Route

To comprehensively outline our approach, we've expressed the logic of our project using a workflow Figure1. Initially, we fine-tuned BART and BART-PubMed on the entire training dataset and compared their overall performance on the test dataset using metrics from three aspects: relevance, readability and factuality. Subsequently, we investigated the performance of the BART-PubMed model across various topics by applying it to subsets of the dataset stratified by keywords. To further enhance performance, we implemented terminology replacement with definitions retrieved from a medical dictionary and evaluated its impact on model performance.

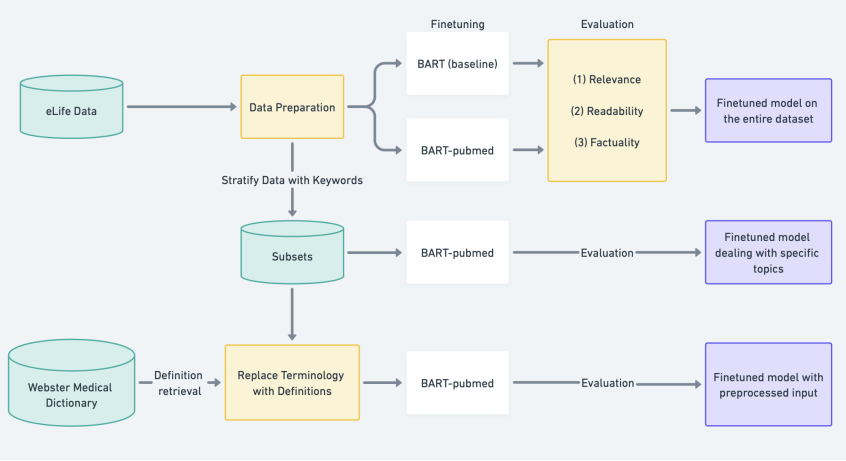

Figure 1: Project workflow diagram

4.2 Models

BART
As discussed previously, BART (Bidirectional and Auto-Regressive Transformers) has proven to be an efficient text generation model. In this section, we delve into the details of the BART model employed in our study. BART is a denoising autoencoder built upon a sequence-to-sequence architecture, rendering it applicable to a wide range of end tasks(12). It employs a standard Transformer-based neural machine translation architecture(20), comprising a bidirectional encoder and an auto-regressive decoder.

During the pre-training stage, BART is trained by corrupting documents and then optimizing the cross-entropy loss between the decoder's output and the original document. This pre-training approach enables BART to learn a robust representation of language, which is effective across a diverse range of text-based tasks.

By leveraging the strengths of both bidirectional encoding and auto-regressive decoding, BART establishes a strong foundation for fine-tuning on specific downstream tasks. Given that our task aims to generate lay summaries of biomedical publications, we employed the BART model fine-tuned on the XSum dataset(21), which consists of news articles with highly abstractive summaries.

BART-PubMed
Since our dataset involves articles from the biomedical domain, we selected another BART-based model that is pretrained on the PubMed dataset(22). In contrast to the previously utilized XSum news dataset, the PubMed dataset encompasses longer articles, enabling the model to be trained on extracting and summarizing more extensive texts. As the PubMed dataset comprises biomedical articles and abstracts, it presents an opportunity to use the articles as inputs and the abstracts as ground-truth labels for supervised learning of neural models. The model utilized 119,924 data for the training set, 6,633 for the validation set, and 6,658 for the test set. This BART-PubMed model not only has a greater ability to summarize lengthy articles but also renders it more suitable for information extraction tasks from biomedical literature, aligning with the domain-specific nature of our dataset.

Transfer Learning
To better adapt the models to our eLife biomedical publications dataset, we employed transfer learning. Transfer learning enables us to transfer the knowledge acquired by a model previously trained on the XSum and PubMed to a new model, tackling similar text summarization tasks. Moreover, it reduces the training time complexity, as the models have already undergone some degree of pre-training(23). Consequently, we fine-tuned the BART model and BART-PubMed model on our eLife dataset and compared their performance. Furthermore, we fine-tuned the BART model and BART-PubMed model on the five topics with the highest article counts within our dataset to evaluate their performance across different topics.

4.3 Evaluation

To comprehensively evaluate our model, results will be evaluated based on three perspectives: (1) relevance, (2) readability, and (3) factuality. Therefore, we select the following evaluation metrics corresponding to the three perspectives in Table3. ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is a package that allows automatic measures to determine the quality of a summary by comparing it to other ideal human-written summaries(24). Specifically, ROUGE-1 and ROUGE-2 belong to the ROUGE-N metrics, which refer to the overlap of unigrams and bigrams between the candidate summaries and the reference summaries. ROUGE-L is based on the Longest Common Subsequence (LCS) to evaluate sentence-level structure similarity. BERTscore computes a similarity score for each token in the candidate sentence with each token in the reference sentence using contextual embeddings(25). As for readability evaluation,

we want to highlight three standard metrics: Flesch-Kincaid grade level (FKGL)(26), Coleman-Liau index (CLI)(27) and Dale–Chall readability score (DCRS)(28) which have the following formula:

Flesch-Kincaid grade level

$$0.39\big(\frac{\text{total words}}{\text{total sentences}}\big) + 11.8\big(\frac{\text{total syllables}}{\text{total words}}\big) - 15.59,$$

Coleman-Liau index

$$0.0588L - 0.296S + 15.8,$$

Dale–Chall readability score

$$0.1579\big(\frac{\text{difficult words}}{\text{total words}} \times 100\big) + 0.0496\big(\frac{\text{total words}}{\text{total sentences}}\big),$$

where L is the average number of letters per 100 words and S is the average number of sentences per 100 words. These readability metrics estimate the year of education needed to understand the generated texts, and lower scores mean higher readability. Finally, the factuality metrics SummaC(29), is more recently introduced to detect inconsistency in summarization. It is designed to assess the consistency between generated summaries and the original documents, a higher SummaC score indicates higher consistency.

Table 3: Evaluation Metrics

	Evaluation Metrics
Relevance	ROUGE (1, 2, and L) and BERTScore
Readability	FKGL, CLI, DCRS
Factuality	SummaC

4.4 Prepocessing

Finally, we aimed to enhance our model's performance through more robust data preprocessing. We replaced medical terminology with definitions from Webster's Medical Dictionary, a comprehensive reference for medical terms. We first deployed Named-Entity Recognition (NER) to identify entities in the articles. If an entity was in the dictionary, we fetched the corresponding definition and substituted the entity with its definition. If it was not found in the dictionary, we left it the original form. The modified dataset will then be evaluated with BART-PubMed, comparing results before and after the terminology replacement.

5 Results

5.1 BART vs BART-PubMed on the Entire eLife Data

In the first experiment, we compared the baseline capability of Biomedical Lay Summarization for BART versus BART-PubMed. BART was found to generate summaries that are more relevant to the target summaries; BART-PubMed excelled in terms of all readability metrics, where lower readability scores indicate higher readability (Table3). Considering that our primary objective in this project is to generate more readable summaries for non-healthcare professionals, we proceeded with BART-PubMed for further experiments. Table4

5.2 BART-PubMed Performance on Stratified Datasets

We next finetuned BART-PubMed separately on the subsets as listed in the Table2. The performance varied across different branches of topics, as shown in Table5 and Table6.

Table 4: BART and BART-PubMed Performance on the Entire eLife Data

Metric	ROUGE-1	ROUGE-2	ROUGE-L	BERT-score	FKGL	CLI	DCRS
BART	**57.82**	**22.44**	**54.16**	**86.93**	9.17	10.62	8.97
BART-PubMed	54.92	19.13	51.06	86.07	**8.71**	**10.13**	**8.71**

The model finetuned on structural biology and molecular biophysics tend to show better relevance scores. In developmental biology, we observed more readable and factual texts, indicated by the FKGL, CLI, and SummaC scores. Besides comparing the absolute values of these evaluation metrics, we computed the relative improvement of readability scores compared to the original data. The model fine-tuned on developmental biology demonstrated the most significant improvement, with readability scores increasing by 4.42% for FKGL, 5.51% for CLI, and 1.81% for DCRS, respectively.

Table 5: BART-PubMed Performance on Domains (Relevance%)

Domain	ROUGE-1	ROUGE-2	ROUGE-L	BERT-score
Biochemistry and Chemical Biology	52.62	16.54	49.15	85.57
Cell Biology	54.27	17.99	50.36	**86.13**
Developmental Biology	54.30	18.88	50.38	86.08
Microbiology and Infectious Disease	54.42	18.83	50.63	86.08
Neuroscience	55.40	**19.85**	51.57	86.05
Structural Biology and Molecular Biophysics	**56.98**	19.36	**51.75**	86.00

Table 6: BART-PubMed Performance on Domains (Readability and Factuality)

Domain	FKGL	CLI	DCRS	SummaC
Biochemistry and Chemical Biology	7.886	9.608	8.869	0.417
Cell Biology	8.863	10.473	8.825	0.417
Developmental Biology	**8.233**	**9.324**	8.530	**0.450**
Microbiology and Infectious Disease	8.762	10.846	8.762	0.395
Neuroscience	8.658	10.360	**8.503**	0.431
Structural Biology and Molecular Biophysics	9.493	10.130	9.185	0.395

5.3 Extra Preprocessing on Training Data

As mentioned in Section 4.4, we retrieved background knowledge from medical dictionary to enrich the information in training data. We did this experiment on the subset with the topic of “Structural biology and molecular biophysics". The results comparing with or without extra preprocessing are illustrated in Table7, where we observe that almost all the metrics gain improved except for BERT-score and SummaC.

6 Discussion and Conclusion

The objective of our project is to generate more readable summaries from biomedical research papers, thereby contributing to knowledge translation and effective communication within the scientific community. Our research results have highlighted the strength of BART-PubMed in generating more readable summaries and the potential for further improvement with an additional definition retrieval step in preprocessing. The limitation is that we only tested definition retrieval on one subset of data. Moving forward, we intend to apply the definition retrieval step across all subsets to ensure

Table 7: Performance Comparing Without vs With Extra Preprocessing

	ROUGE-1	ROUGE-2	ROUGE-L	Bert-score	FKGL	CLI	DCRS	SummaC
w/o	53.805	16.223	49.575	85.433	8.667	9.924	8.851	41.403
with	**58.871**	**20.084**	**56.669**	84.144	**8.137**	**9.618**	**8.071**	40.955

consistency in the results. We fine-tuned BART-PubMed separately for different branches of topics to investigate its baseline capabilities in summarizing these topics. These results may inspire future work to design topic-specialized biomedical lay summarization solutions. Additionally, recognizing the challenge posed by limited input length, another direction to expand our experiments is to include the Longformer Encoder Decoder (LED) model. LED, based on BART, is tailored to handle longer sequences. Based on our approach in this project and following these strategic steps, it is promising that we can foster greater accessibility to biomedical research.

7 Contribution Statement

We here outline the contributions of each team member towards the completion of this project:

- Experiment - BART vs BART-PubMed on the entire training data: Xincheng Cai
- Experiment - BART and BART-PubMed on six sub-datasets: Mengmeng Du
- Experiment - NER, terminology replacement, and test: Mengmeng Du
- Evaluation Pipeline: Mengmeng Du and Xincheng Cai
- Project Presentation: Mengmeng Du and Xincheng Cai
- Report - Introduction, Dataset, Methods: Mengmeng Du
- Report - Abstract, Related Work, Results, Discussion: Xincheng Cai
- Github Repository: Mengmeng Du and Xincheng Cai

We extend our sincere gratitude to Prof. Arman Cohan for his guidance and insightful feedback, which significantly contributed to the refinement of our experimental plan.

References

[1] T. Goldsack, Z. Luo, Q. Xie, C. Scarton, M. Shardlow, S. Ananiadou, and C. Lin, "Biolaysumm 2023 shared task: Lay summarisation of biomedical research articles," in *Workshop on Biomedical Natural Language Processing*, 2023.

[2] C.-H. Chiang, Y.-S. Chuang, and H.-Y. Lee, "Recent advances in pre-trained language models: Why do they work and how do they work," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pp. 8–15, 2022.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[5] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[11] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[12] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. rahman Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Annual Meeting of the Association for Computational Linguistics*, 2019.

[13] B. Wang, Q. Xie, J. Pei, Z. Chen, P. Tiwari, Z. Li, and J. Fu, "Pre-trained language models in biomedical domain: A systematic survey," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–52, 2023.

[14] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[15] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[16] K. Huang, J. Altosaar, and R. Ranganath, "Clinicalbert: Modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019.

[17] K. Huang, A. Singh, S. Chen, E. T. Moseley, C.-Y. Deng, N. George, and C. Lindvall, "Clinical xlnet: modeling sequential clinical notes and predicting prolonged mechanical ventilation," *arXiv preprint arXiv:1912.11975*, 2019.

[18] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, and S. Yu, "Biobart: Pretraining and evaluation of a biomedical generative language model," *arXiv preprint arXiv:2204.03905*, 2022.

[19] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "Biogpt: generative pre-trained transformer for biomedical text generation and mining," *Briefings in bioinformatics*, vol. 23, no. 6, p. bbac409, 2022.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[21] S. Narayan, S. B. Cohen, and M. Lapata, "Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization," 2018.

[22] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, "A discourse-aware attention model for abstractive summarization of long documents," 2018.

[23] G. Vrbančič and V. Podgorelec, "Transfer learning with adaptive fine-tuning," *IEEE Access*, vol. 8, pp. 196197–196211, 2020.

[24] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, pp. 74–81, 2004.

[25] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," 2020.

[26] J. Kincaid, R. Fishburne, R. Rogers, and B. Chissom, "Flesch-kincaid grade level," *Memphis: United States Navy*, 1975.

[27] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring.," *Journal of Applied Psychology*, vol. 60, no. 2, p. 283, 1975.

[28] J. Chall and E. Dale, *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, 1995.

[29] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, "Summac: Re-visiting nli-based models for inconsistency detection in summarization," 2021.

Reproducibility checklist

* Please make sure these points are addressed in your report submission
* Please copy this and replace the ☐ with a ✔ for the items that are addressed in your report/code submission
* Please complete this report, attach it to your final project report as the last page and then submit.

Model Description, algorithm, Mathematical Setting:

- ✔ Include a thorough explanation of the model/approach or the mathematical framework

Source Code Accessibility:

- ✔ Provide a link to the source code on github.
- ✔ Ensure the code is well-documented
- ✔ Ensure that the github repo has instructions for setting up the experimental environment.
- ✔ Clearly list all dependencies and external libraries used, along with their versions.

Computing Infrastructure:

- ✔ Detail the computing environment, including hardware (GPUs, CPUs) and software (operating system, machine learning frameworks) specifications used for your results.
    - (Example statement 1: the model was fine-tuned using a single T4 GPU on colab.
      Example statement 2: we ran inference of Llama 70B using 4 Nvidia A5000 GPUs)
- ✔ Mention any specific configurations or optimizations used.
    - (Example: We used a quantized version of Llama with int8.
      Example 2: We used the regular float32 representation.)

Dataset Description:

- ✔ Clearly describe the datasets used, including sources, preprocessing steps, and any modifications.
- ✔ If possible, provide links to the datasets or instructions on how to obtain them.

Hyperparameters and Tuning Process:

- ✔ Detail the hyperparameters used and the process for selecting them.
    - (Example: The model was fine-tuned using a batch size of 16, learning rate of 1e-5, and trained on 1000 steps with 100 steps of learning rate linear warmup with linear decay)

Evaluation Metrics and Statistical Methods:

- ✔ Clearly define the evaluation metrics and statistical methods used in assessing the model.

Experimental Results:

- ✔ Present a comprehensive set of results, including performance on test sets and/or any relevant validation sets.
- ✔ Include comparisons with baseline models and state-of-the-art, where applicable.

Limitations and future work:

- ✔ Include a discussion of the limitations of your approach and potential areas for future work.