Lay Summarization: Bridging Biomedical Research and Everyday Understanding

Chloe Nguyen
University of California, Berkeley
chloenguyen@berkeley.edu

Chih Ling (Phoebe) Yueh
University of California, Berkeley
phoebeyueh0401@berkeley.edu

Abstract

Understanding biomedical research papers can be difficult for non-experts, including patients, caregivers, and policymakers, due to their specialized technical content. To improve accessibility, this study investigates generating lay summaries that present complex biomedical articles in general terms. We comprehensively explore various abstractive and hybrid summarization techniques, leveraging models such as BART, BERT, and domain-specific adaptations. Our approach evaluates accuracy and readability to ensure that the generated summaries effectively convey the core concepts of the research articles while remaining accessible to a general audience. Using the Public Library of Science (PLOS) dataset, we conduct experiments with different summarization strategies, including fine-tuning models on article subsets and combining extractive and abstractive methods. We evaluate the results using accuracy metrics, through ROUGE, and readability metrics, through the Flesch-Kincaid Grade Level and Gunning Fog Index. Our findings demonstrate that the abstract provides sufficient information for many abstractive and hybrid techniques (fine-tuned on essential and non-technical subsections or all of the articles) to be applied onto and generate summaries that are both more accurate and more readable. This suggests a promising approach for enhancing the accessibility of biomedical research papers to a general audience.

1 Introduction

Biomedical research papers often contain specialized information that can be challenging for non-experts (i.e., a lay person), including patients, caregivers, and policymakers, to understand. As it creates a barrier to the dissemination and practical application of scientific findings, we aim to create lay summaries that present the core content of these papers in general terms, making them more accessible to address this issue. In this paper, we comprehensively explore various abstractive and hybrid summarization techniques to generate lay summaries from biomedical articles, leveraging models such as BART, BERT, and domain-specific adaptations. Our approach focuses on accuracy and readability to ensure that the generated summaries effectively convey the essence of the research studies.

2 Background

Summarization tasks are typically categorized as either extractive or abstractive. For extractive summarization, earlier approaches like TextRank (Mihalcea and Tarau, 2004) and LexRank (Erkan and Radev, 2011) apply graph-based algorithms to rank sentences. Building upon the foundation, a more modern approach, BERTSUM proposed by Liu (2019), utilizes BERT embeddings for sentence ranking. For abstractive summarization, there have been significant advancements in the field with various model developments. Pointer-Generator Networks, introduced by See et al. (2017), allow the model to "point" to the most relevant source text and still produce novel sentences. T5 (Text-to-Text Transfer Transformer) by Raffel et al. (2019) uses a unified approach and treats all the text processing problems as "text-to-text" generation. The well-known BART, developed by Lewis et al. (2019), is a denoising autoencoder for pretraining sequence-to-sequence models. PEGASUS by Zhang et al. (2019), designed specifically for summarization tasks, is a Transformer Encoder-Decoder based model that pretrains by masking and predicting important sen-

tences in a document. However, previous studies have also identified significant difficulties in abstractive summarization, notably the generation of factual inconsistencies and hallucinations, where the model introduces content that is not present in the source text (Goldsack et al., 2022). It has also been observed that state-of-the-art models often default to extractive summarization, copying parts of the original article rather than generating novel summaries (Subramanian et al., 2019).

There has been little experimentation in long-text lay summarization within particular domains such as biomedicine due to several challenges, including the availability of data sources, the complexity of biomedical texts, and the architectural limitations of existing language models. CL-LaySumm 2020, one of the shared tasks at Scholarly Document Processing, is considered as the most well-known and attracted eight submissions. The top 1 system designed by Chandrasekaran et al. (2020) used PEGASUS to produce abstractive summaries and Presumm (Liu and Lapata, 2019) to add to existing summaries if the word count is below threshold. Summaformers fed different combinations of article sections as inputs and experimented with both fine-tuned BART and T5 (Roy et al., 2021). Inspired by Roy et al. (2021), we also want to experiment with limiting the model inputs to subsets of the article that contain essential non-technical sections. Rather than fine-tuning on the full article, restricting the information fed into the model could potentially yield better performance.

Another aspect we aim to address differently from other studies is the examination of readability scores alongside accuracy. Guo et al. (2020) emphasized that removing unnecessary scientific jargon and simplifying sentence structures are crucial for successful lay summarization, so they proposed to use both automated metrics (e.g., Flesch-Kincaid grade level, Gunning Fog index, and Coleman-Liau index) and human evaluation to assess readability. Due to resource constraints, we will only utilize two of the three automated metrics in the paper.

3 Dataset

We used the Public Library of Science (PLOS) subset of the Scientific Lay Summarisation dataset, which contains long biomedical articles paired with non-technical summaries. The dataset is divided into 24,773 training, 1,376 validating and 1,376 testing sets. The version available on Hugging Face consolidates each article's sections into a single string labeled as 'article', so we preprocess the 'article' column by segmenting each section using "/n", which we believe is the optimal way for extracting inputs in subsequent model training stages.

4 Methodology

4.1 Baseline

To establish a starting point for our experiments, we implemented two baseline approaches, leveraging abstractive summarization with the BART model pre-trained on CNN/DailyMail dataset. The first baseline performed abstractive summarization on the abstract of each research paper. The second baseline applied abstractive summarization to the abstract, the first paragraph of the introduction, and the conclusion. Our subsetting approach was motivated by See et al. (2017)'s findings, which suggested that models performed better when applied to this specific subset of the articles.

4.2 Proposed Approaches

Our study aims to generate lay summaries from research articles, which involves the challenge of simplifying complex and technical information while retaining the core ideas of the research. Inspired by past research, we aim to focus the model on specific subsections of the paper to avoid overwhelming it with technical intricacies that are unnecessary for a lay summarization. Previous studies have examined the application of models on different subsections- Roy et al. (2021) found high performance when applying BART on the abstract, while See et al. (2017) found that the combination of the abstract, introduction, and conclusion performed best. Taking into account accuracy and readability, we apply these subsetting methods as we propose variations of two main techniques: abstractive summarization and hybrid summarization.

4.2.1 Abstractive Summarization Approaches

To address readability and prevent generated summaries from being overwhelmed with technical details, our abstractive summarization experiments test limiting information to the model to

an article subset of essential, non-technical sections. This is done in two ways: fine-tuning the model on a subsection of the articles and/or feeding the model subsections of the article. The motivation behind this approach is to focus the model on less technical sections that should already contain the core ideas of the research paper - specifically the abstract, introduction, and conclusion. We hypothesize that fine-tuning the model on these sections, rather than the full article, will promote the generation of more easily understandable summaries. This approach is based on the idea that these sections provide sufficient context while avoiding overly technical details.

The proposed abstractive summarization approaches are:

- **Experiment: Abstracts w/ BART Fine-Tuned on Abs, Intro & Con**
  This approach applied BART fine-tuned on the abstract, introduction and conclusion on the abstracts of each article.

- **Experiment: Abstracts w/ BART Fine-Tuned on Full Article**
  This approach applied BART fine-tuned on the full article on the abstracts of each article.

- **Experiment: Lead-K w/ BART Fine-Tuned on Abs, Intro & Con**
  This approach applied BART fine-tuned on the full article on the first K sentences of the abstracts of each article, where K is the average summary length. Lead-K has been proven to work particularly well with news articles (See et al., 2017), as articles are typically structured to prioritize the most critical information in the beginning.

- **Experiment: Abs, Intro & Con w/ BART Fine-Tuned on Abs, Intro & Con**
  This approach applied BART fine-tuned on the abstract, introduction and conclusion on the abstract, introduction and conclusion of each article.

4.2.2 Hybrid Summarization Approaches

To handle simplifying complex and technical information while retaining the core ideas of the research, our hybrid summarization experiments focus on identifying (via extractive summarization) and rephrasing (via abstractive summarization) the most important sentences to enhance the quality and readability of the summaries. The motivation behind this approach was because many human-generated lay summaries include reworded sentences corresponding to sentences in the full technical paper- suggested from the findings of See et al. (2017).

We proposed to compare different extractive methods to select the most relevant information to feed into the abstractive summarization model- specifically, we examine a non-domain-specific approach using TF-IDF and domain-specific deep learning model approach using Bio-BERT. Additionally, we combined this with our proposed abstractive summarization experiments to limit information to the model and explore the performance of our abstractive BART model fine-tuned on a concise subset of the articles (abstract, introduction, and conclusion) versus the full articles.

The proposed hybrid summarization approaches are:

- **Experiment: Abstracts w/ TF-IDF & BART Fine-Tuned on Abs, Intro & Con**
  This approach applied TF-IDF to perform extractive summarization on the abstract, then uses BART (fine-tuned on the abstract, introduction and conclusion) to generate summaries.

- **Experiment: Abstracts w/ TF-IDF & BART Fine-Tuned on Full Article**
  This approach applied TF-IDF to perform extractive summarization on the abstract, then uses BART (fine-tuned on the full articles) to generate summaries.

- **Experiment: Abstracts w/ TF-IDF & Pre-Trained CNN BART**
  This approach applied TF-IDF to perform extractive summarization on the abstract, then uses BART (pre-trained on CNN data) to generate summaries.

- **Experiment: Abs, Intro & Con w/ TF-IDF & BART Fine-Tuned on Abs, Intro & Con**
  This approach applied TF-IDF to perform extractive summarization on the abstract, introduction and conclusion of each article, then uses BART (fine-tuned on the abstract, introduction and conclusion) to generate summaries.

- **Experiment: Abs, Intro & Con w/ TF-IDF & BART Fine-Tuned on Full Article**

This approach applied TF-IDF to perform extractive summarization on the abstract, introduction and conclusion of each article, then uses BART (fine-tuned on the full articles) to generate summaries.

- **Experiment: Abstracts w/ Bio-BERT & BART Fine-Tuned on Abs, Intro & Con**
  This approach applied Bio-BERT to perform extractive summarization on the abstract of each article, then uses BART (fine-tuned on the abstract, introduction and conclusion) to generate summaries.

- **Experiment: Abstracts w/ Bio-BERT & BART Fine-Tuned on Full Article**
  This approach applied Bio-BERT to perform extractive summarization on the abstract of each article, then uses BART (fine-tuned on the full articles) to generate summaries.

- **Experiment: Abs, Intro & Con w/ Bio-BERT & BART Fine-Tuned on Abs, Intro & Con**
  This approach applied Bio-BERT to perform extractive summarization on the abstract, introduction and conclusion of each article, then uses BART (fine-tuned on the abstract, introduction and conclusion) to generate summaries

- **Experiment: Abs, Intro & Con w/ Bio-BERT & BART Fine-Tuned on Full Article**
  This approach applied Bio-BERT to perform extractive summarization on the abstract, introduction and conclusion of each article, then uses BART (fine-tuned on the abstract, introduction and conclusion) to generate summaries.

4.3 Measuring Success

We evaluated our results on two metric categories: accuracy and readability. By considering both factors, our study aims to generate summaries that capture core information of a research article while still being comprehensible to a lay audience.

4.3.1 Accuracy Metrics

Our accuracy metrics aim to assess how effectively essential information from research articles is captured. This is evaluated using ROUGE scores to compare the generated summaries against the human-generated lay summaries. High ROUGE scores suggest that the generated summaries accurately reflect the core content of the original research articles.

4.3.2 Readability Metrics

Our readability metrics aim to assess how readable and understandable the summaries are to the general public. This is evaluated through two scores:

- Flesch-Kincaid grade: Estimates the educational level required to understand the text. Lower scores suggest that the summary is easier to understand.

- Gunning Fog Index: Measures the complexity of the text, considering both sentence length and word difficulty. Lower scores indicate greater readability and accessibility.

4.4 Fine-Tuning Process

To set up the abstractive summarization process of our experiments, we fine-tuned four BART models using the training subset of PLOS with various training configurations and evaluated the models based on their training and validation losses over multiple epochs.

We began with fine-tuning the models on a subset of the paper: the abstract, the first paragraph of the introduction, and the conclusion of each paper. Model 1 was trained on the specified subset and showed steadily declining training losses - however, the validation losses still remain quite high. Since these results suggest that the model is still learning, there is potential for further improvement with more training time. With these observations, we increased the learning rate and extended the training epochs to 4 in Model 2. The higher learning rate helped the model converge faster and the training loss was lower than Model 1, but there wasn't a substantial validation loss reduction. The training plot for this model showed that the training rate and duration were appropriate (refer to Appendix - Figure 1). To further verify the best number of epochs, we tried 10-epochs fine-tuning in Model 3. The experiment confirmed that, with the current learning rate, 3-4 epochs allowed for the lowest validation loss and larger epochs lead to overfitting, as validation losses increase. From these explorations, we determined Model 2 to demonstrate the best performance.

We also implemented fine-tuning on the full article to examine model performance when given

increased article context. Thus, Model 4 was trained on the full articles, instead of using only three paragraphs. The results showed a lower training loss compared to previous models, with a comparable final validation loss to Model 1 and 2. The training plot for this model showed a stable decline curve, suggesting that we had either reached or were close to the minimum loss and so, we did not see a need to further fine-tune the hyperparameters (refer to Appendix - Figure 2).

5 Results and Discussion

Examining the results in Table 2, we see that the best overall performing model in our experiments was the abstractive summarization model Abstracts w/ BART Fine-Tuned on Abs, Intro & Con, where abstracts were fed into a BART model that was fine-tuned on the abstract, introduction, and conclusion sections. The best performing hybrid summarization model was Abstracts w/ Bio-BERT & BART Fine-Tuned on Abs, Intro & Con (Model 2 Run), where abstracts were fed into Bio-BERT and then into a BART model that was fine-tuned on abstract, introduction, and conclusion sections. Both of these models demonstrated significantly better ROUGE accuracy scores than the baseline models- with the best abstractive summarization experiment showing slightly higher accuracy scores across all ROUGE metrics than the best hybrid summarization experiment. Additionally, their readability scores were closer to those of the reference human-generated lay summaries - although interestingly, the readability scores were higher, indicating that these experiment-generated summaries were less readable than the baseline generated summaries.

Moreover, all abstractive summarization experiments demonstrated better performance than the baseline models in terms of both accuracy and readability, with clear improvement in ROUGE accuracy scores and with readability scores that are more similar to that of the reference human-generated summaries. On the other hand, hybrid summarization experiments demonstrated mixed performance compared to the baseline models. There was a notable difference in performance depending on the extractive summarization method used. Experiments using Bio-BERT extractive summarization consistently showed better accuracy and slightly higher readability scores than the TF-IDF method. This is understandable because Bio-BERT is pre-trained on relevant domain knowledge, allowing it to better identify important sentences when applied to biomedical research articles in comparison to a non-domain specific frequency-based metric like TF-IDF. As such, while the best Bio-BERT experiments showed significantly better accuracy than the baseline and produced readability scores closer to those of the reference human-generated lay summaries, TF-IDF experiments consistently performed worse in accuracy compared to the baseline and produced readability scores roughly equal to the baseline.

Another trend to note was that the readability scores of baseline models with BART pre-trained on CNN baselines were consistently lower than the readability scores of experiments with BART fine-tuned on the research articles. This is likely because the language used in research articles is generally more complex than that used in CNN news. As such, a model trained on research articles is likely to learn more technical language.

When examining results of various approaches to limit the information that the model is performed on, we consistently observed improved accuracy when applying the fine-tuned model to only the abstract in comparison to applying the fine-tuned model to the abstract, introduction, and conclusion subsections. This may be because focusing on the abstract alone allows the model to concentrate on the most relevant information and adding extra context from the introduction and conclusion provides unnecessary information for the model when it comes to generating lay summaries. Readability scores between just abstract and abstract, introduction, and conclusion experiments were roughly similar. This is expected, as the average language complexity of the abstract alone should be comparable to that of the combined sections. Moreover, when examining results of various approaches to limit the information the model is fine-tuned on, performance of experiments trained on just the abstracts, introductions and conclusions were similar to that of experiments trained on the full articles for accuracy and readability scores.

Overall, our ROUGE accuracy scores are comparable to those in existing literature, such as the BART-based approaches in See et al. (2017), Roy et al. (2021), and Goldsack et al. (2022)- with our best models performing in competitive accuracy metric ranges: 0.45-0.5 range for ROUGE-1

Model	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1	Flesch Kincaid	Gunning Fog Index
Reference Lay Summaries	-	-	-	14.7387	16.2056
Baselines					
Abstracts w/ BART Pre-Trained on CNN	0.306763	0.113582	0.195037	13.1357	15.7926
Abs, Intro & Con w/ BART Pre-Trained on CNN	0.352977	0.125774	0.212946	13.3373	15.7965

Table 1: Reference Lay Summaries & Baseline Models

Model	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1	Flesch Kincaid	Gunning Fog Index
Abstractive Summarization					
Abstracts w/ BART Fine-Tuned on Abs, Intro & Con (Model 2)	**0.473110**	**0.166371**	**0.256382**	14.476	16.4794
Abstracts w/ BART Fine-Tuned on Full Article (Model 4)	0.471058	0.164585	0.254511	14.459	16.4701
Lead-8 w/ BART Fine-Tuned on Abs, Intro & Con (Model 2)	0.457128	0.158671	0.248594	14.440	16.5462
Abs, Intro & Con w/ BART Fine-Tuned on Abs, Intro & Con (Model 2)	0.460008	0.152612	0.242909	14.4701	16.6164
Hybrid Summarization					
Abstracts w/ TF-IDF & BART Fine-Tuned on Abs, Intro & Con (Model 2)	0.255242	0.071991	0.161721	13.5278	15.1336
Abstracts w/ TF-IDF & BART Fine-Tuned on Full Article (Model 4)	0.276755	0.073226	0.168256	14.0819	15.5774
Abstracts w/ TF-IDF & Pre-Trained CNN BART	0.224291	0.064978	0.149229	10.725	12.2482
Abs, Intro & Con w/ TF-IDF & BART Fine-Tuned on Abs, Intro & Con (Model 2)	0.254575	0.071607	0.161537	13.4978	15.1234
Abs, Intro & Con w/ TF-IDF & BART Fine-Tuned on Full Article (Model 4)	0.276792	0.073070	0.168045	14.1371	15.6169
Abstracts w/ Bio-BERT & BART Fine-Tuned on Abs, Intro & Con (Model 2)	0.450988	**0.156606**	0.241409	14.8397	16.9232
Abstracts w/ Bio-BERT & BART Fine-Tuned on Full Article (Model 4)	**0.451763**	0.156527	**0.242421**	14.5994	16.6332
Abs, Intro & Con w/ Bio-BERT & BART Fine-Tuned on Abs, Intro & Con (Model 2)	0.301541	0.055709	0.153247	12.8773	15.2849
Abs, Intro & Con w/ Bio-BERT & BART Fine-Tuned on Full Article (Model 4)	0.295603	0.056633	0.154623	12.3036	14.6189

Table 2: Fine-tuned Models

F1, 0.15-0.2 range for ROUGE-2 F1 and 0.25-0.3 range for ROUGE-L F1.

6 Limitations

While our approach aims to bridge the gap between complex biomedical papers and lay understanding, we acknowledged several limitations within the process.

- **Dataset Constraints:** Because lay summarization is less pervasive compared to other text summarization NLP tasks, the availability of large training datasets is limited (Goldsack et al., 2022). Hence, the relatively smaller training dataset size of 24,773 articles in the PLOS dataset may not be sufficient to achieve high accuracy and robustness in training effective models. Moreover, it limits the generalizability of the model to other domains, as biomedical texts often contain terminologies, jargon, and complexities specific to the field.

- **Model Limitations:** The BART model has its inherent limitations regarding the maximum input token lengths, which is typically capped at 1024 tokens. As biomedical articles often exceed this length by a significant amount, truncating them to fit within the model's input limit may result in the omission of critical information.

- **Readability Challenges:** ROUGE metrics, while standard in summarization tasks, may not fully capture the quality of lay summaries in terms of comprehensibility for a general audience. Although we incorporate readability metrics like Flesch-Kincaid Grade Level and Gunning Fog Index to further address this problem, they do not account for the additional complexity introduced by medical abbreviations (Guo et al., 2020). Conducting human evaluations is also resource-intensive, posing a challenge for large-scale assessment and iterative model improvement.

7 Conclusion

Our research tackled the challenge of making biomedical research papers more accessible to non-experts by generating lay summaries, examining accuracy and readability scores. We successfully proposed several abstractive and hybrid summarization approaches to focus the model on relevant information, generating lay summaries with improved accuracy and desirable readability scores over a baseline of BART pre-trained on CNN news. In particular, our results suggest that abstractive summarization models fine-tuned on the essential, non-technical sections of articles (specifically, the abstract, introduction and conclusion sections), when applied solely on the research abstracts, demonstrated particularly promising results in both accuracy and readability.

References

Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard H. Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer, and Anita de Waard. 2020. Overview and insights from the shared tasks at scholarly document processing 2020: Cl-scisumm, laysumm and longsumm. In *SDP*.

Günes Erkan and Dragomir R. Radev. 2011. Lexrank: Graph-based lexical centrality as salience in text summarization. *CoRR*, abs/1109.2128.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. *CoRR*, abs/2210.09932.

Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2020. Automated lay language summarization of biomedical scientific reviews. *CoRR*, abs/2012.12573.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *CoRR*, abs/1908.08345.

Yang Liu. 2019. Fine-tune BERT for extractive summarization. *CoRR*, abs/1903.10318.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In Dekang Lin and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Sayar Ghosh Roy, Nikhil Pinnaparaju, Risubh Jain, Manish Gupta, and Vasudeva Varma. 2021. Summaformers @ laysumm 20, longsumm 20. *CoRR*, abs/2101.03553.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368.

Sandeep Subramanian, Raymond Li, Jonathan Pilault, and Christopher J. Pal. 2019. On extractive and abstractive neural document summarization with transformer language models. *CoRR*, abs/1909.03186.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777.

Appendix

Model	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1	Flesch Kincaid	Gunning Fog Index
Abstractive Summarization					
PEGASUS (pretrained on CNN dailymail)	25.8688	9.4590	16.9448	12.7281	15.2926
PEGASUS (pretrained on PubMed)	27.8301	5.6920	16.4429	14.1769	13.7785
Hybrid Summarization					
Abs, Intro & Con w/ TF-IDF & BART Fine-Tuned on Abs, Intro & Con (Model 1)	25.2079	7.0594	16.0278	13.4366	15.0511
Abs, Intro & Con w/ TFIDF & BART Fine-Tuned on Abs, Intro & Con (Model 3)	25.9544	6.9583	16.0087	13.8536	15.7788

Table 3: Extra Model Results

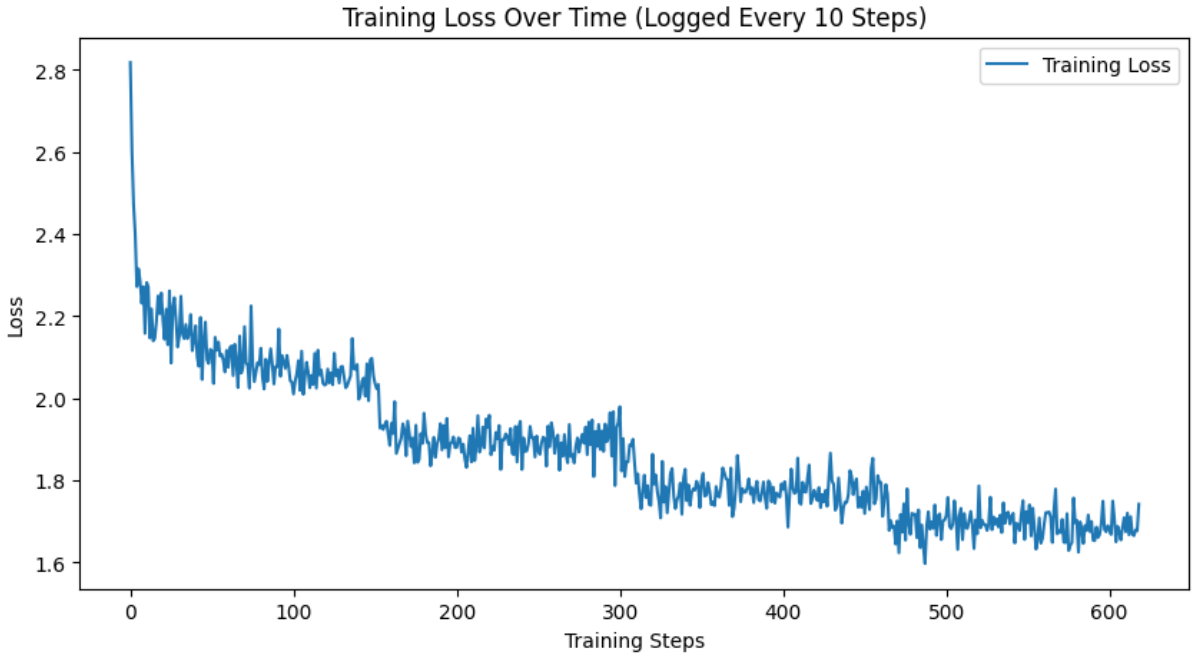

Figure 1: Training plot for Model 2

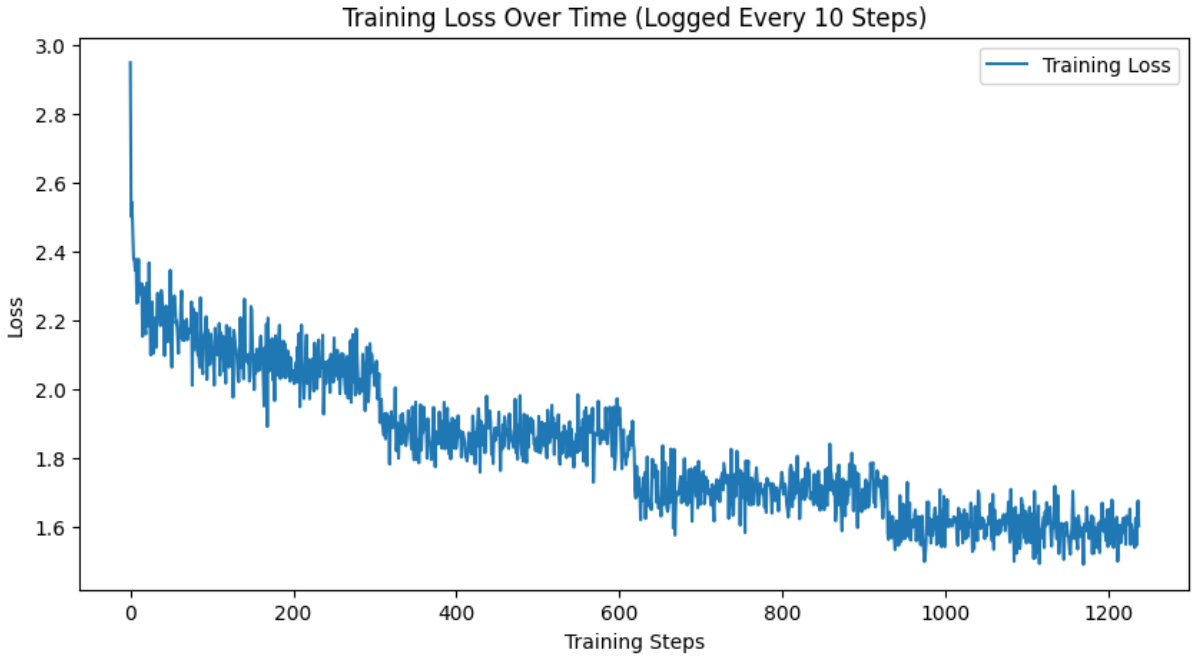

Figure 2: Training plot for Model 4