Adaptive BPE Tokenization for Enhanced Vocabulary Adaptation in Finetuning Pretrained Language Models

Gunjan Balde[§], Soumyadeep Roy[§], Mainack Mondal and Niloy Ganguly
Indian Institute of Technology Kharagpur
balde.gunjan0812@kgpian.iitkgp.ac.in
soumyadeep.roy9@iitkgp.ac.in
{mainack,niloy}@cse.iitkgp.ac.in

Abstract

In this work, we show a fundamental limitation in vocabulary adaptation approaches that use Byte-Pair Encoding (BPE) tokenization scheme for fine-tuning pretrained language models (PLMs) to expert domains. Current approaches trivially append the target domain-specific vocabulary (V_{DOMAIN}) at the end of the PLM vocabulary. This approach leads to a lower priority score and causes sub-optimal tokenization in BPE that iteratively uses merge rules to tokenize a given text. To mitigate this issue, we propose ADAPTBPE where the BPE tokenization initialization phase is modified to first perform the longest string matching on the added (target) vocabulary before tokenizing at the character level. We perform an extensive evaluation of ADAPTBPE versus the standard BPE over various classification and summarization tasks; ADAPTBPE improves by 3.57% (in terms of accuracy) and 1.87% (in terms of Rouge-L), respectively. ADAPTBPE for MEDVOC works particularly well when reference summaries have high OOV concentration or are longer in length. We also conduct a human evaluation, revealing that ADAPTBPE generates more relevant and more faithful summaries as compared to MEDVOC. We make our codebase publicly available at https://github.com/gb-kgp/adaptbpe.

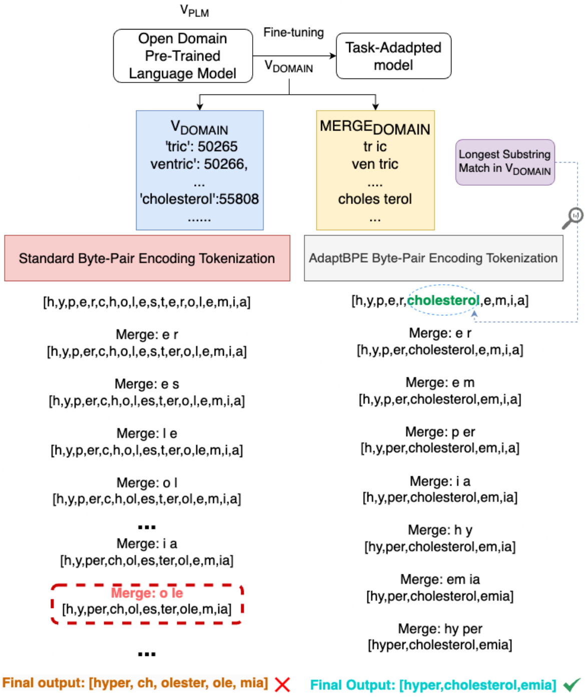

V_{PLM}
Open Domain Pre-Trained Language Model
Fine-tuning
V_{DOMAIN}
Task-Adadpted model
V_{DOMAIN}
'tric': 50265
ventric': 50266,
...
'cholesterol':55808
......
$MERGE_{DOMAIN}$
tr ic
ven tric
....
choles terol
...
Longest Substring Match in V_{DOMAIN}
Standard Byte-Pair Encoding Tokenization
AdaptBPE Byte-Pair Encoding Tokenization
[h,y,p,e,r,c,h,o,l,e,s,t,e,r,o,l,e,m,i,a]
Merge: e r
[h,y,p,er,c,h,o,l,e,s,t,er,o,l,e,m,i,a]
Merge: e s
[h,y,p,er,c,h,o,l,es,t,er,o,l,e,m,i,a]
Merge: l e
[h,y,p,er,c,h,o,l,es,t,er,o,le,m,i,a]
Merge: o l
[h,y,p,er,c,h,ol,es,t,er,ol,e,m,i,a]
...
Merge: i a
[h,y,per,ch,ol,es,ter,ol,e,m,ia]
Merge: o le
[h,y,per,ch,ol,es,ter,ole,m,ia]
...
Final output: [hyper, ch, olester, ole, mia] ✗
[h,y,p,e,r,cholesterol,e,m,i,a]
Merge: e r
[h,y,p,er,cholesterol,e,m,i,a]
Merge: e m
[h,y,p,er,cholesterol,em,i,a]
Merge: p er
[h,y,per,cholesterol,em,i,a]
Merge: i a
[h,y,per,cholesterol,em,ia]
Merge: h y
[hy,per,cholesterol,em,ia]
Merge: em ia
[hy,per,cholesterol,emia]
Merge: hy per
[hyper,cholesterol,emia]
Final Output: [hyper,cholesterol,emia] ✓

Figure 1: ADAPTBPE modifies the initialization step of standard BPE by merging the characters that match with the extended vocabulary (V_{DOMAIN}). The incorrect merge step of BPE for tokenizing the word *hypercholesterolemia* is highlighted by a red dashed box.

1 Introduction

Vocabulary adaptation-based fine-tuning has proved successful in domain adaptation to expert domains, characterized by high vocabulary mismatch. Here, the PLM vocabulary is further extended by adding a target domain-specific vocabulary (V_{DOMAIN}) during fine-tuning. To identify V_{DOMAIN} works like VOLT (Xu et al., 2021) and AVOCADO (Hong et al., 2021) focus on optimizing the model's vocabulary by adding subwords based on utility scoring functions that are based on variants of fragment score (Rust et al., 2021) or optimize Pointwise Mutual Information (Diao et al., 2021) or by measuring domain shift of token sequence distribution (Sachidananda et al., 2021). MEDVOC (Balde et al., 2024), is the first work in a summarization setting that uses fragment score as the utility function. **In this work, we establish the need also to adapt the tokenization scheme**; Figure 1 provides an example of ill-tokenization due to the limitations of the standard BPE tokenization scheme.

Prior vocabulary adaptation studies (Hong et al., 2021; Balde et al., 2024) append added vocabulary and corresponding merge rules towards the end of

[§] Both the authors have contributed equally to this work. Email id of Corresponding author: balde.gunjan0812@kgpian.iitkgp.ac.in.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 14724–14733
November 12-16, 2024 ©2024 Association for Computational Linguistics

existing PLM vocabulary ($\mathrm{V_{PLM}}$). **This approach does not guarantee that the Byte-Pair Encoding (BPE) tokenizer will use the added target domain vocabulary**. We believe this is because the merge rules are trivially appended to the end, automatically implying a lower priority (of $\mathrm{V_{DOMAIN}}$) over existing PLM vocabulary ($\mathrm{V_{PLM}}$).

Our main contribution is to propose the ADAPTBPE tokenization scheme that mitigates the above-mentioned ill-tokenization issue of BPE when applied to vocabulary adaptation strategies. Our proposed ADAPTBPE algorithm is independent of the target domain-specific vocabulary construction algorithm and only modifies the underlying BPE tokenization phase. **ADAPTBPE modifies the initialization stage of a standard BPE tokenization** as explained in detail in Algorithm 1. Instead of starting tokenization by splitting the input token to character level, ADAPTBPE performs the longest substring match in the added vocabulary ($\mathrm{V_{DOMAIN}}$) iteratively and preserves the matched substring from splitting into characters further. This modified BPE algorithm, ADAPTBPE, mitigates the ill-tokenization issues completely, as we observe a significant drop in fragment score (average number of subwords a given word across the entire corpus) of 39.16% and 13.96% in case of AVOCADO and MEDVOC respectively.

ADAPTBPE shows improvements of 3.57% and 1.87% over the standard BPE algorithm in the case of AVOCADO and MEDVOC respectively for eight datasets (4 classification and 4 summarization tasks). In the case of MEDVOC for difficult generation scenarios such as high OOV (out-of-vocabulary) concentration and longer reference summaries, ADAPTBPE consistently improves by 10.41% and 3.30% in terms of Rouge-L. We further perform a human evaluation using medical experts where we observe that ADAPTBPE produces more relevant and faithful summaries in the case of MEDVOC. We make our codebase publicly available at https://github.com/gb-kgp/adaptbpe.

2 Background

Vocabulary Adaptation Strategies for Classification –AVOCADO. AVOCADO (Hong et al., 2021) propose a vocabulary adaptation strategy for classification tasks. AVOCADO iteratively adds task-specific vocabulary ($\mathrm{V_{DOMAIN}}$) constructed from source documents of target tasks to existing PLM vocabulary ($\mathrm{V_{PLM}}$). The amount of vocabulary to be added is decided using fragment score, which is defined as the average number of subwords tokenized per word given a vocabulary. AVOCADO starts on a set of words that are split into more than two subwords ($W_{s>2}$) and constructs task-specific vocabulary on this set of words. It then keeps on adding the vocabulary from this task-specific vocabulary till the fragment score of words in set ($W_{s>2}$) stays above a fixed threshold, γ. AVOCADO initialized the embeddings of the newly added subwords with the average of embeddings of the subwords they were previously split into. AVOCADO uses contrastive loss framework (Chen et al., 2020) as a regularization loss along with the standard cross-entropy loss for classification to tune the model with the added embeddings of the newly added subwords.

Vocabulary Adaptation Strategies for Summarization –MEDVOC. MEDVOC (Balde et al., 2024) proposes a vocabulary adaptation framework for summarization tasks in the medical domain for three models – BERT, BART, and PEGASUS. First, MEDVOC identifies vocabulary to be added ($\mathrm{V_{DOMAIN}}$) as an optimizable parameter. It constructs vocabulary on candidate set of medical OOV (Out-Of-Vocabulary) words (words that are medical, and split into more than one word using existing PLM vocabulary) identified from combination of PAC (PubMed Abstract Collection) dataset ($\mathrm{V_{PAC}}$) and target-task specific datasets ($\mathrm{V_{TGT}}$). It then performs a hyperparameter search using fragment score as the metric, over different vocabulary sizes and identifies the optimal vocabulary to be added to existing PLM vocabulary ($\mathrm{V_{PLM}}$). The embeddings are initialized randomly and are tuned by performing an intermediate fine-tuning step on PAC dataset comprising PubMed abstract as source document and the title as reference summary.

3 Proposed Methodology

Working of the standard BPE Tokenization. BPE is the most common tokenization scheme that is found to be most effective among various tokenization strategies (Gallé, 2019; Zouhar et al., 2023; Schmidt et al., 2024), and is used in the majority of recent Large Language Models (LLMs) like LLaMa (Touvron et al., 2023a,b) and Mistral (Jiang et al., 2023). BPE tokenization scheme takes as input two files: (i) vocabulary file, which contains the vocabulary, and (ii) merge rules file, which contains merge rules for the terms present in the vocabulary required for its construction (e.g., *th e* merge rule for the word *'the'* in vocabulary). The standard BPE tokenizer starts by splitting the input word into the character level. Then, following a bottom-up strategy, it iteratively merges adjacent characters following the ordered merge rules from the merge rule file taken as input. For instance, consider the word *happy*. BPE starts by converting this word into a list of characters: *[h, a, p, p, y]*. Then, it checks for possible merges on adjacent characters and selects the one with the least rank. Here, it chooses <*p,p*> resulting in *[h, a, pp, y]*. It then iteratively keeps checking and ends with [*'happy'*] as the final output for BPE tokenization.

ADAPTBPE Tokenization Scheme (Algorithm 1). We observe that the main reason for ill-tokenization (See Figure 1) is certain merge rules that hinder the formation of added vocabulary. Therefore, instead of splitting at the character level at the initialization stage, we first check for the longest substring match (Hofmann et al., 2022) only in the added vocabulary (V_{DOMAIN}) and prevent the match from splitting into the character level. This step is iterated till we cannot find any substring match. Figure 1 shows an example: the word *hypercholesterolemia* is initialized as *[h,y,p,e,r,cholesterol,e,m,i,a]* as opposed to standard BPE tokenization which starts entirely at character level: *[h,y,p,e,r,c,h,o,l,e,s,t,e,r,o,l,e,m,i,a]*.

4 Experimental Setup

We use the same experimental setup as the state-of-the-art vocabulary adaptation works of AVOCADO and MEDVOC for the classification and summarization tasks, respectively. Appendix A provides all the necessary implementation details.

Datasets. We use the same datasets as used in AVOCADO and MEDVOC (see Appendix A.2 for further details) — (i) four classification tasks: CHEMPROT (Kringelum et al., 2016) from the biomedical domain, ACL-ARC (Jurgens et al., 2018) from the computer science domain, HYPERPARTISAN (HYP) (Kiesel et al., 2019) from the news domain, and AMAZON (McAuley et al., 2015) from the customer reviews domain, and (ii) two query-focused document summarization datasets: EBM (Mollá and Santiago-Martínez, 2011) and BioASQ (Tsatsaronis et al., 2015), and two question summarization datasets: MeQSum (Ben Abacha and Demner-Fushman, 2019) and CHQSum (Yadav et al., 2022).

Algorithm 1: ADAPTBPE tokenization

```
Input: Text text, Tokenizer tokenizer, Merge rules merges, Added
       vocabulary $V_{DOMAIN}$
Output: BPE token sequence $\mathcal{T}$
// Pre-tokenizing the text based on pre-tokenization rules of
   tokenizer
pre_tokenized ← tokenizer.pre_tokenize_str(text)
pre_tokenized_text ← [word for word in pre_tokenized]
$\mathcal{T}$ ← []
for word ∈ pre_tokenized_text do
    split ← {}
    // Finding the longest substring match in $V_{DOMAIN}$
    remaining ← word
    while $True$ do
        $idx_{match}$ , $longest_{match}$ ← longest_substr(remaining,
          $V_{DOMAIN}$)
        if $idx_{match} == -1$ then
            break
        else
            split[$idx_{match}$] ← $longest_{match}$
            for i in range($idx_{match}$,
              $idx_{match}$+$longest_{match}$.length) do
                remaining[i] ← '-'
    // Retrieving the longest matches and remaining parts of
       string
    subwords ← [sw for i, sw in sorted(split)]
    pairs ← get_bigrams(subwords)
    // Standard BPE loop
    while $True$ do
        bigram ← {least ranking applicable merge rule on pairs}
        if bigram is invalid then
            break
        first, second ← bigram
        new_word ← []
        i := 0
        while $i <$ subwords.length do
            $j$ := subwords.index(first, i)
            new_word.extend(subwords[i:j])
            $i := j$
            if subwords[i] == first and i < len(subwords) - 1
              and subwords[i + 1] == second then
                new_word.append(first + second)
                $i := i + 2$
            else
                new_word.append(subwords[i])
                $i := i + 1$
        new_word := tuple(new_word)
        subwords := new_word
        if subwords.length == 1 then
            break
        else
            pairs ← get_bigrams(subwords)
    $\mathcal{T}$.extend(subwords)
return $\mathcal{T}$
```

Evaluation Metrics. We report classification accuracy and Macro-F1 scores for the classification task. For summarization, we report Rouge-L (Lin, 2004) and Concept Score (Zhang et al., 2023), which measures the overlap of UMLS medical concepts between the generated and reference summaries. See Appendix A.3 for additional details.

5 Performance Evaluation of ADAPTBPE

We show the performance comparison results of ADAPTBPE in Table 1 and 2 for the vocabulary adaptation strategies for the classification and summarization setting, respectively. Please refer Ap-

Dataset (Domain)	**Model**	**FragSr**	**Accuracy**	**Macro-F1**
CHEMPROT	BPE	2.55	**81.43** ± 0.55	54.88 ± 1.66
(BioMedical)	ADAPTBPE	**1.16**	81.40 ± 0.40	**55.02** ± 0.47
ACL-ARC*	BPE	2.21	69.03 ± 5.05	55.04 ± 8.24
(Computer Science)	ADAPTBPE	**1.18**	**73.02** ± 4.21	**62.00** ± 4.95
HYP	BPE	3.26	77.84 ± 5.20	74.23 ± 7.01
(News)	ADAPTBPE	**3.17**	**82.16** ± 2.50	**80.64** ± 3.03
AMAZON	BPE	2.81	83.13 ± 3.64	68.34 ± 0.47
(Reviews)	ADAPTBPE	**2.47**	**86.26** ± 0.53	**69.90** ± 0.29

Table 1: Performance evaluation of ADAPTBPE on AVOCADO with RoBERTa-Base as base model averaged across 5 seeds (* -except for ACL-ARC which was done for 20 seeds). Improvements wherever observed are statistically significant (t-test: p-value$<$ 0.05). We show improvements of 3.57% in accuracy and 3.18% in the case of the Macro-F1 score. ADAPTBPE results in the fragment score (FragSr) drop of 39.16% across datasets.

pendix A.4 for the sizes of added vocabularies in both the settings for all the datasets and Appendix A.5 for relevant hyperparameter details.

Performance Evaluation in Classification Datasets. We show the performance comparison of BPE versus ADAPTBPE in Table 1. We observe gains of 3.57% in accuracy and 3.18% in the case of the Macro-F1 score. We further observe huge drops in fragment scores of 39.16% across four datasets from four domains. Thus, ADAPTBPE helps to correctly tokenize the domain words, which leads to better performance.

Performance Evaluation in Medical Summarization Datasets. We show the performance comparison of BPE versus ADAPTBPEin Table 2. We observe gains of 1.87% in Rouge-L (R-L) and 0.9% in the case of ConceptScore (CSr). We further observe huge drops in fragment scores of 13.20% across four datasets. This indicates the efficacy of ADAPTBPE, as we are now correctly tokenizing the words and thus the downstream task improvement. We further investigate how ADAPTBPE performed compared to BPE when reference summaries had high OOV concentration and were long in length following evaluation as performed in MEDVOC. These points represent the most difficult data points in terms of vocabulary mismatch. ADAPTBPE shows a big improvement of 10.40% on average across four datasets in high OOV settings and 3.41% on average across BioASQ and EBM datasets for a long-form generation.

Human Evaluation. We randomly select 40 test data points sampled uniformly from four summarization datasets and follow the annotation procedure as described in (Fabbri et al., 2021; Balde et al., 2024) to get annotations of summaries across the dimensions of *relevance*, *coherence* (on a Likert scale of 1 to 5), and *faithfulness* (binary). Each annotator was given 30 minutes to evaluate 10 summaries and was compensated at a rate of 8 UK pounds per hour (see Appendix B for more details). Figure 2 shows the human evaluation results where ADAPTBPE generates more faithful summaries (97.5% vs. 77.5% of summaries are faithful), and more relevant summaries, where 82.5% of data points get a positive score (≥ 4) in Likert scale, as compared to 65% in case of BPE for MEDVOC.

Model	**FragSr**	**R-L$_{\text{All}}$**	**CSr$_{\text{All}}$**	**R-L$_{\text{H-O}}$**	**R-L$_{\text{L-RS}}$**
			EBM		
BPE	3.00	20.65	22.66	19.23	17.62
ADAPTBPE	**2.31**	**20.73**	**22.67**	**21.43**	**17.74**
			BioASQ		
BPE	3.14	48.02	52.87	39.23	43.25
ADAPTBPE	**2.71**	47.72	**52.93**	**42.95**	**45.91**
			MeQSum		
BPE	3.34	55.88	60.52	75.56	-
ADAPTBPE	**3.15**	**58.00**	**62.29**	**82.64**	-
			CHQ		
BPE	2.94	40.59	**45.63**	33.77	-
ADAPTBPE	**2.67**	**41.92**	44.57	**37.60**	-

Table 2: Performance evaluation of ADAPTBPE on MEDVOC model with BART-Large as base model. We observe gains of 1.87% in Rouge-L (**R-L**) and 0.9% in Concept Score (**CSr**). Improvements wherever observed are statistically significant (t-test: p-value$<$ 0.001). In High-OOV settings (R-L$_{\text{H-O}}$) we observe gains of 10.40% and 3.30% in long-form generation (**R-L$_{\text{L-RS}}$** considering only EBM and BioASQ). Notably, ADAPTBPE results in the fragment score (FragSr) drop of 13.20% across datasets.

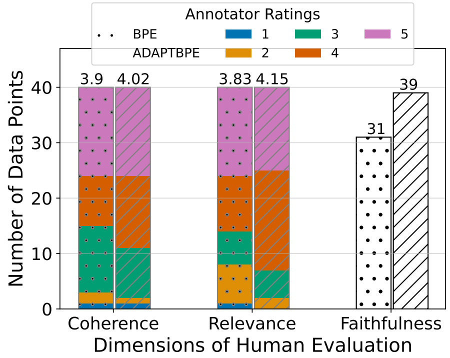


Figure 2: Human evaluation scores comparison over 40 randomly selected test data points. ADAPTBPE produces more relevant, coherent, and faithful summaries during human evaluation with medical experts.

6 Conclusion

We are the first to show the incorrect BPE tokenization issue present in vocabulary adaptation techniques for fine-tuning PLMs to the target (expert) domain, designed for both classification and summarization tasks from various domains. The newly added target domain vocabulary is trivially added at low priority, causing BPE tokenizers to ignore them. Therefore, we propose a novel BPE tokenization scheme, ADAPTBPE, that modifies the BPE initialization step by searching through V_{DOMAIN} to find the longest substring match. Our proposed ADAPTBPE algorithm is independent of the target domain-specific vocabulary construction algorithm and focus only on improving the tokenization part. ADAPTBPE-enabled models outperform the competing baselines by 3.57% and 1.87% on average over classification and summarization tasks, respectively. Human evaluation using medical experts rate ADAPTBPE-based summaries to be more relevant and faithful than standard BPE.

7 Limitations

We limit our evaluation to only pretrained language models and do not show results on large language models that also utilize BPE, such as LLaMa or Mistral, which uses Sentencepiece (Kudo and Richardson, 2018) Byte-level BPE Tokenization with fallback. We observe that 27.76% of target domain-specific vocabulary terms are still tokenized into more than one subword (i.e., the ill-tokenization issue persists) for MEDVOC in the case of the LlaMa-2-7B model. However, the models considered in this study (BART and RoBERTa) use *huggingface tokenizers* library (Wolf et al., 2020) and we observed ill-tokenization in 64.13% of target domain-specific vocabulary terms. Thus, some efforts are needed to make ADAPTBPE work for LLMs. Second, the issue of ill-tokenization is mostly prevalent in the case of BPE but less prevalent in the case of WordPiece tokenization, which is used by BERT and does not exist for Unigram tokenization scheme, which is used by PEGASUS and FLAN-T5 models.

8 Ethics Statement and Broader Impact

Summarization and other text generation systems powered by large language models can suffer from hallucinations, producing outputs that deviate from the source material and are unfaithful summaries. While the proposed ADAPTBPE tokenization scheme generates more faithful summaries compared to existing baselines based on human evaluation, the summaries from such AI models are not yet reliable enough for high-stakes applications like medical contexts involving professionals and clinicians. Substantially more research is still needed to understand better the types of faithfulness and relevance errors made by these AI systems and to ultimately develop methods to mitigate or prevent such errors before these technologies can be safely deployed in sensitive real-world settings.

References

Gunjan Balde, Soumyadeep Roy, et al. 2024. Medvoc: Vocabulary adaptation for fine-tuning pre-trained language models on medical text summarization. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6180–6188.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234.

Ting Chen, Simon Kornblith, et al. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Shizhe Diao, Ruijia Xu, et al. 2021. Taming pre-trained language models with n-gram representations for low-resource domain adaptation. 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 3336–3349.

Alexander R. Fabbri, Wojciech Kryściński, et al. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Matthias Gallé. 2019. Investigating the effectiveness of BPE: The power of shorter sequences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1375–1381.

Valentin Hofmann, Hinrich Schuetze, et al. 2022. An embarrassingly simple method to mitigate undesirable properties of pretrained language model tokenizers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 385–393.

Jimin Hong, TaeHee Kim, et al. 2021. AVocaDo: Strategy for adapting vocabulary to downstream domain. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4692–4700.

Albert Q. Jiang, Alexandre Sablayrolles, et al. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

David Jurgens, Srijan Kumar, et al. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Johannes Kiesel, Maria Mestre, et al. 2019. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839.

Jens Kringelum, Sonny Kim Kjaerulff, et al. 2016. Chemprot-3.0: a global chemical biology diseases mapping. *Database*, 2016:bav123.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, et al. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Yinhan Liu, Myle Ott, et al. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Julian McAuley, Christopher Targett, et al. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 43–52.

Diego Mollá and María Elena Santiago-Martínez. 2011. Development of a corpus for evidence based medicine summarisation. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 86–94.

Phillip Rust, Jonas Pfeiffer, et al. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. 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 3118–3135.

Vin Sachidananda, Jason Kessler, and Yi-An Lai. 2021. Efficient domain adaptation of language models via adaptive tokenization. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 155–165, Virtual. Association for Computational Linguistics.

Craig W. Schmidt, Varshini Reddy, et al. 2024. Tokenization is more than compression. *Preprint*, arXiv:2402.18376.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, SIGIR*, pages 1–4.

Hugo Touvron, Thibaut Lavril, et al. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

George Tsatsaronis, Georgios Balikas, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Thomas Wolf, Lysandre Debut, et al. 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.

Yonghui Wu, Mike Schuster, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *Preprint*, arXiv:1609.08144.

Jingjing Xu, Hao Zhou, et al. 2021. Vocabulary learning via optimal transport for neural machine translation. 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 7361–7373.

Shweta Yadav, Deepak Gupta, et al. 2022. Chq-summ: A dataset for consumer healthcare question summarization. *arXiv preprint arXiv:2206.06581*.

Nan Zhang, Yusen Zhang, et al. 2023. FaMeSumm: Investigating and improving faithfulness of medical summarization. pages 10915–10931.

Vilém Zouhar, Clara Meister, et al. 2023. A formal perspective on byte-pair encoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 598–614.

A Experimental Setup

A.1 Pre-trained Language Models

To test the generalizability of our method described in Section 3, we evaluate the efficacy of MEDVOC on BART in case of summarization and RoBERTa in case of classification.

- **RoBERTa** (Liu et al., 2019): RoBERTa (Robustly Optimized BERT Approach) is a transformer-based model, enhancing the original BERT model by training with more data and improved training techniques. It eliminates the Next Sentence Prediction (NSP) task used in BERT and employs dynamic masking during pre-training to increase robustness. RoBERTa is trained on a diverse corpus, including the Common Crawl dataset, to better capture nuanced language patterns. This model achieves state-of-the-art performance on various natural language processing (NLP) benchmarks. We use RoBERTa-base[§] which is a 125 Million parameter model and uses Byte-pair Encoding tokenization case with a vocabulary ($|\mathrm{V}_{\mathrm{PLM}}|$) of size 50265.

- **BART** (Lewis et al., 2020): BART is a denoising autoencoder, implemented as a sequence-to-sequence model with a bidirectional encoder over corrupted text and a left-to-right auto-regressive decoder to generate the original document it was derived from. We use the BART-LARGE[§] model available from the *huggingface* library. BART has 406 Million parameters, uses *Byte-Pair Encoding* tokenization, and its pretraining objective is a combination of *Text Infilling* and *Sentence Shuffling*. The vocabulary size of this PLM ($|\mathrm{V}_{\mathrm{PLM}}|$) is 50265.

A.2 Datasets

We describe here the details on the target task dataset mentioned briefly in Section 4.

Classification We use four target task datasets for classification that were used in AVoCaDo. The dataset stats are described in Table 3.

- **CHEMPROT** (Kringelum et al., 2016). Chemprot dataset is a corpus used for the task of chemical-protein relation extraction. It consists of scientific abstracts annotated with various types of interactions between chemical compounds and proteins, such as inhibition, activation, and binding in total 13 classes. The dataset is commonly used to train and evaluate models in the domain of biomedical natural language processing, particularly for the extraction and classification of biochemical relationships.

- **ACL-ARC** (Jurgens et al., 2018). The ACL-ARC dataset is designed to classify the intent behind citations in academic papers. It consists of annotated citations from research papers in the ACL Anthology, categorizing each citation based on its purpose, such as background, use, or comparison in total 6 classes. This dataset aids in understanding the functional and rhetorical roles of citations in scholarly communication.

- **HYPERPARTISAN** (Kiesel et al., 2019). The hyperpartisan dataset consists of news articles labeled for hyperpartisanship, indicating whether they exhibit extreme bias. It was created to support research in detecting biased or partisan news content –a two-class classification, providing annotations on article-level and publisher-level partisanship. This dataset is used in natural language processing tasks to develop and benchmark models for identifying and understanding media bias.

- **AMAZON** (McAuley et al., 2015). The amazon dataset is a comprehensive collection of customer reviews and ratings from Amazon, covering a wide range of products. It includes detailed reviews, ratings, product metadata, and user information, providing a rich resource for sentiment analysis, recommendation systems, and other NLP tasks. The task is to identify whether a given review as input is actually helpful or not.

Domain	Dataset	Document count			Classes	OOV %
		Train	Val	Test		RoBERTa
BIOMED	CHEMPROT	4169	2427	3469	relation (13)	21.65
CS	ACL-ARC	1688	114	139	citation intent (6)	12.56
NEWS	HYPERPARTISAN	515	65	65	partisanship (2)	3.94
REVIEWS	AMAZON	115251	5000	25000	helpfulness (2)	3.69

Table 3: Dataset statistics of downstream classification datasets. *OOV%* refers to the median fraction of unigrams in SD that are absent from the PLM vocabulary.

Medical Summarization We use four target task datasets in this study: two query-focussed summarization datasets, EBM and BioASQ, and two recent benchmark medical question summarization datasets, MeQSum and CHQSum, each of which we describe below.

- **EBM** (Mollá and Santiago-Martínez, 2011). Here input to the system is a query along with a PubMed abstract, and the expected output is the summary answering the question with the PubMed Abstract as the context.

- **BioASQ** (Tsatsaronis et al., 2015). We use the dataset from BioASQ-9B Phase-B summarization task. The input to the system is a

[§] https://huggingface.co/FacebookAI/roberta-base

[§] https://huggingface.co/facebook/bart-large question followed by relevant snippets from a collection of PubMed Abstracts. There are two kinds of outputs an exact answer and an ideal answer associated with the input. For the summarization task, we consider the ideal answer as the Reference summary.

- **MeQSum** (Ben Abacha and Demner-Fushman, 2019). The dataset is created for better medical question summarization because the original patients' questions are verbose. The dataset contains 1000 patients' health questions selected from a collection distributed by the U.S. National Library of Medicine. Each question is annotated with a summarized question by medical experts.

- **CHQSum** (Yadav et al., 2022). CHQSum consists of 1507 domain-expert annotated question-summary pairs from the Yahoo community question answering forum[§] which provides community question answering threads containing users' questions on multiple diverse topics and the answers submitted by other users. The authors with the help of 6 domain experts identified valid medical question from the forum and asked the experts to formulate an abstractive summary for the questions.

Dataset	Document count			Word count		OOV %
	Train	Val	Test	SD	RS	BART
EBM	1423	209	424	298	58	11.5
BioASQ	1525	491	496	505	40	9.4
MeQSum	700	150	150	70	12	5.7
CHQSum	1000	107	400	184	12	6.3

Table 4: Dataset statistics of downstream medical summarization datasets. *OOV%* refers to the median fraction of unigrams in RS that are absent from the PLM vocabulary.

A.3 Evaluation Metrics

We first describe the implementation details for computing Rouge scores discussed in Section 4, where we use the official Rouge (Lin, 2004) script[§]. The following parameters: *-c 95 -2 -1 -U -r 1000 -n 4 -w 1.2 -a*, are used and we report the median at a 95% confidence interval. Additionaly, we also use Concept Score which identifies the medical conocept overlaps between generated and reference summary. To identify concepts, we use matcher.match utility of QuickUMLS (Soldaini and Goharian, 2016) tool in default setting.

A.4 Added Vocabulary Sizes

We mention the size of added vocabulary obtained by AVocaDo and MEDVOC on classification and summarization datasets in Table 5.

Dataset	$\|V_{\text{DOMAIN}}\|$
AVocaDo ($\|V_{\text{PLM}}\|$: 50265)	
CHEMPROT	5103
ACL-ARC	3419
AMAZON	1168
HYPERPARTISAN	743
MEDVOC ($\|V_{\text{PLM}}\|$: 50265)	
EBM	11061
BioASQ	6462
MeQSum	747
CHQSum	680

Table 5: Size of added vocabulary ($|V_{\text{DOMAIN}}|$) for AVocaDo(RoBERTa) and MEDVOC(BART) on classification and summarization datasets respectively.

A.5 Hyperparameters

We discuss the following hyperparameters: (i) the training hyperparameters, (iii) inference hyperparameters for MEDVOC.

A.5.1 Training Hyperparameters

AVocaDo. All AVocaDo related experiments were run on one V100 32 GB graphic card. We kept the training hyperparameters same as that of what authors follow in the study. In brief, we tune learning rate : $\in \{1e-5, 2e-5, 5e-5\}$ and temperature: from 1.5 to 3.5 in steps of 0.5.

MEDVOC. All the experiments are run on one A100 40 GB GPU. We use the fine-tuning summarization scripts for BART provided in MEDVOC's codebase. We used the following hyperparameters to train BART model. learning rate: 5e-5, batch size: 32, and gradient accumulation steps: 8, rest all the hyperparameters takes its default values. We checkpoint at every 500 steps and train the model for a total of 5 epochs (approx 15K steps). The training times for IFT-PAC for MEDVOC is mentioned in Table 6.

A.5.2 Inference Hyperparameters

We used beam search to run the **inference** on the test set. We tuned the following hyperparameters of beam search: beam size (B $\in [2, 10]$) and length-penalty (Wu et al., 2016) ($lp \in [0.1, 3]$) on the validation split of the target task dataset. The best values of hyperparameters thus obtained are mentioned in Table 7.

[§] https://webscope.sandbox.yahoo.com/catalog.php?datatype=l&did=11

[§] https://github.com/bheinzerling/pyrouge/tree/master

Dataset	BART
EBM	27 hrs 51 mins
BioASQ	28 hrs 25 mins
MeQSum	28 hrs 49 mins
CHQSum	28 hrs 38 mins

Table 6: Time required in hours for intermediate fine-tuning using PAC for each target task dataset using BART with ADAPTBPE.

Dataset	B	lp
EBM	3	0.8
BioASQ	6	0.8
MeQSum	8	0.1
CHQSum	6	0.5

Table 7: Optimal values for inference hyperparameters - beam size (B) and Length Penalty (lp) used for beam-search generation for each of the datasets using BART with ADAPTBPE.

B Human Evaluation

Twelve individuals took part in an annotation task on the Prolific platform. Each person was asked to annotate ten random pairs of summaries from a pool of forty, with the order and source of the summaries concealed. Participants had thirty minutes to finish the task and were paid 8 UK Pounds per hour for their time. They also provided feedback on the experience and demographic information, excluding any personal details beyond what is made available by the platform. The task was conducted using Google Forms, with participants being shown a consent notice beforehand.

Participation Criteria. The filtering criteria for participants were kept same as that of MED-VOC (Balde et al., 2024):

- **Age:** ≥ 25,
- **Primary Language:** English,
- **Highest education level completed:** Graduate degree (MA/MSc/MPhil/other), Doctorate degree (PhD/other)
- **Subject:** Medicine, Health and Medicine, Biomedical Sciences.

Annotation Guidelines. The annotations were carried across three dimensions (Fabbri et al., 2021) of coherence, relevance, and factual consistency. **Coherence** judges how well formed the summaries are and whether the sentences in the summaries are actually related to each other or not. **Relevance** judges how informative the summaries are considering the input as the context for evaluating relevance. **Factual Consistency** judges whether the facts, figures, numbers stated in the generated summary ca be verified from source input or not. Even if the generated text contains correct fact, but cannot be verified by only looking at input it is deemed as factually incosistent.

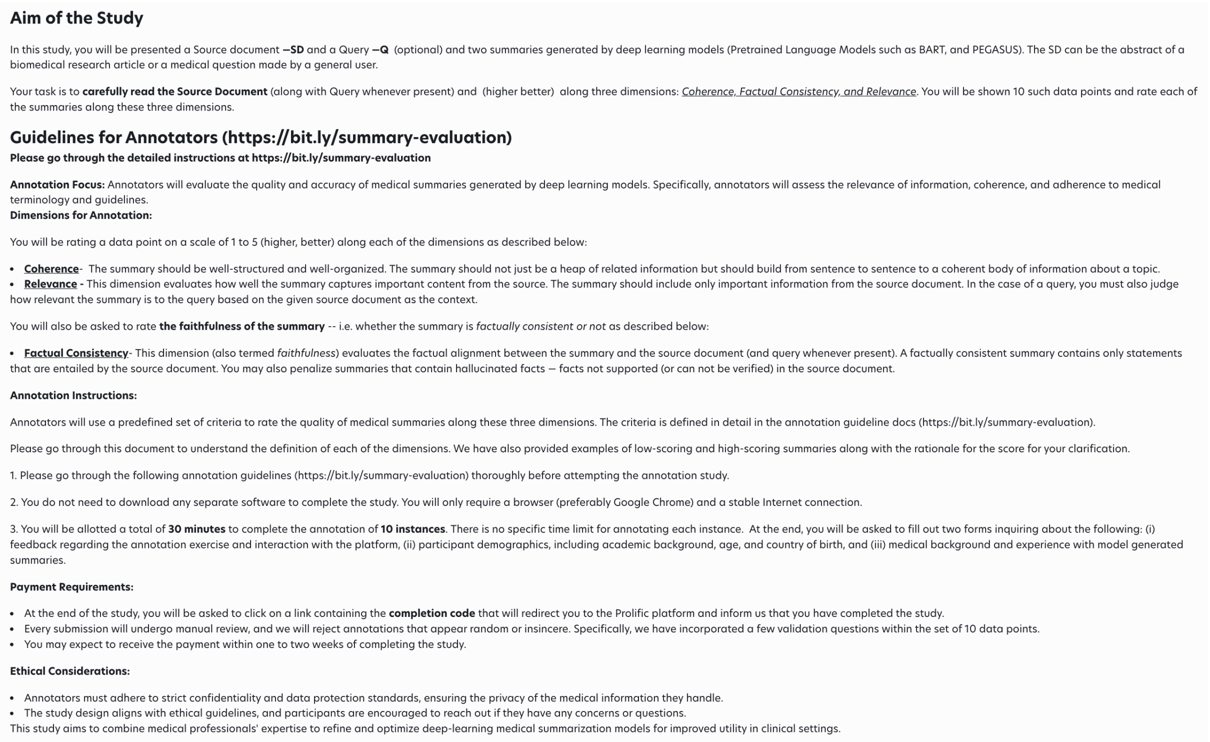
Source Document	GE: Question in laymen terms: Has any genetic or other correlation ever been made between these two diagnosis? My 59 y.o. sister has a diagnosis of Periventricular Heterotopia. Her 30 y.o. daughter has been suffering with same for last 15 years. Her 37 Y.O. daughter is clinically full-care retarded (since infancy) and has severe idiopathic scoliosis. I have severe idiopathic scoliosis. I use the term "severe" to express debilitating and multiple fusion surgeries. All four of my generation female siblings have a level of scoliosis. FYI: this PH sister died last week, her remains are at the [LOCATION]
Positive Example	
Summary	Can there be a genetic link between Periventricular Heterotopia and scoliosis?
Relevance	5
Coherence	5
Factual Consistency	1
Explanation	Here we can see the summary is focused on idenitfying whether a genetic link exists b/w Periventricular Heterotopia and scoliosis which is what the user is asking about.
Negative Example	
Summary	What are the causes of severe idiopathic scoliosis?
Relevance	1
Coherence	1
Factual Consistency	0
Explanation	The question is asking for treatments of scoliosis which is not the theme of the input document.

Table 8: A negative and positive example as shown to the participant in the annotation guidelines for clarification under the three dimensions of annotation. The data point is taken from MeQSum dataset.

For each of these dimensions, we show one positive (high rating) and one negative example (low rating) along with an explanation as a part of our annotation guideline (Table 8).

Demographic analysis of participants. The average age of participants was 29 years. Out of 12 participants, 10 were female and 2% were male. All the participants are Graduate studtents. The participants were recruited by platfrom from 3 countries: South Africa(3), Sweden(2), and UK(7).

Instruction on platform. Prolific begins the user study with a clear instruction window describing what the task is about and what the participant is expected to do in the study. We attach the screenshot of that window which is shown to the participants in Figure 3.



Aim of the Study

In this study, you will be presented a Source document **—SD** and a Query **—Q** (optional) and two summaries generated by deep learning models (Pretrained Language Models such as BART, and PEGASUS). The SD can be the abstract of a biomedical research article or a medical question made by a general user.

Your task is to **carefully read the Source Document** (along with Query whenever present) and (higher better) along three dimensions: *Coherence, Factual Consistency, and Relevance*. You will be shown 10 such data points and rate each of the summaries along these three dimensions.

Guidelines for Annotators (https://bit.ly/summary-evaluation)

Please go through the detailed instructions at https://bit.ly/summary-evaluation

Annotation Focus: Annotators will evaluate the quality and accuracy of medical summaries generated by deep learning models. Specifically, annotators will assess the relevance of information, coherence, and adherence to medical terminology and guidelines.
Dimensions for Annotation:

You will be rating a data point on a scale of 1 to 5 (higher, better) along each of the dimensions as described below:

- **Coherence**- The summary should be well-structured and well-organized. The summary should not just be a heap of related information but should build from sentence to sentence to a coherent body of information about a topic.
- **Relevance -** This dimension evaluates how well the summary captures important content from the source. The summary should include only important information from the source document. In the case of a query, you must also judge how relevant the summary is to the query based on the given source document as the context.

You will also be asked to rate **the faithfulness of the summary** -- i.e. whether the summary is *factually consistent or not* as described below:

- **Factual Consistency**- This dimension (also termed *faithfulness*) evaluates the factual alignment between the summary and the source document (and query whenever present). A factually consistent summary contains only statements that are entailed by the source document. You may also penalize summaries that contain hallucinated facts — facts not supported (or can not be verified) in the source document.

Annotation Instructions:

Annotators will use a predefined set of criteria to rate the quality of medical summaries along these three dimensions. The criteria is defined in detail in the annotation guideline docs (https://bit.ly/summary-evaluation).

Please go through this document to understand the definition of each of the dimensions. We have also provided examples of low-scoring and high-scoring summaries along with the rationale for the score for your clarification.

1. Please go through the following annotation guidelines (https://bit.ly/summary-evaluation) thoroughly before attempting the annotation study.

2. You do not need to download any separate software to complete the study. You will only require a browser (preferably Google Chrome) and a stable Internet connection.

3. You will be allotted a total of **30 minutes** to complete the annotation of **10 instances**. There is no specific time limit for annotating each instance. At the end, you will be asked to fill out two forms inquiring about the following: (i) feedback regarding the annotation exercise and interaction with the platform, (ii) participant demographics, including academic background, age, and country of birth, and (iii) medical background and experience with model generated summaries.

Payment Requirements:

- At the end of the study, you will be asked to click on a link containing the **completion code** that will redirect you to the Prolific platform and inform us that you have completed the study.
- Every submission will undergo manual review, and we will reject annotations that appear random or insincere. Specifically, we have incorporated a few validation questions within the set of 10 data points.
- You may expect to receive the payment within one to two weeks of completing the study.

Ethical Considerations:

- Annotators must adhere to strict confidentiality and data protection standards, ensuring the privacy of the medical information they handle.
- The study design aligns with ethical guidelines, and participants are encouraged to reach out if they have any concerns or questions.

This study aims to combine medical professionals' expertise to refine and optimize deep-learning medical summarization models for improved utility in clinical settings.

Figure 3: Instruction window as seen by an annotator participating in the study.