Evaluation of LLMs in Medical Text Summarization: The Role of Vocabulary Adaptation in High OOV Settings

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

Large Language Models (LLMs) recently achieved great success in medical text summarization by simply using in-context learning. However, these recent efforts do not perform fine-grained evaluations under difficult settings where LLMs might fail. They typically report performance scores over the entire dataset. Through our benchmarking study, we show that LLMs show a significant performance drop for data points with high concentration of out-of-vocabulary (OOV) words or with high novelty. Vocabulary adaptation is an intuitive solution to this vocabulary mismatch issue where the LLM vocabulary gets updated with certain expert domain (here, medical) words or subwords. An interesting finding from our study is that Llama-3.1, even with a vocabulary size of around 128K tokens, still faces *over-fragmentation* issue with medical words. To that end, we show vocabulary adaptation helps improve the LLM summarization performance even in difficult settings. Through extensive experimentation of multiple vocabulary adaptation strategies, two continual pretraining strategies, and three benchmark medical summarization datasets, we gain valuable insights into the role of vocabulary adaptation strategies for customizing LLMs to the medical domain. We also performed a human evaluation study with medical experts where they found that vocabulary adaptation results in more relevant and faithful summaries. Our codebase is made publicly available at https://github.com/gb-kgp/LLM-MedicalSummarization-Benchmark.

1 Introduction

Recent works like Clinsumm (Van Veen et al., 2024) explore various strategies to adapt LLMs in the task of medical text summarization. These strategies are (i) in-context learning (Brown et al., 2020; Lampinen et al., 2022) where a fixed number of exemplars is added in the prompt at inference time only; and (ii) parameter-efficient finetuning using QLoRA (Dettmers et al., 2024) because LLMs have billions of parameters which make complete finetuning computationally infeasible. However, these studies do not perform fine-grained evaluations considering challenging generation scenarios and typically report performance scores over the entire dataset. In this paper, we consider two such challenging scenarios primarily related to the high vocabulary mismatch between LLM vocabulary and the medical domain. Figure 1 provides an overview of our benchmarking setup.

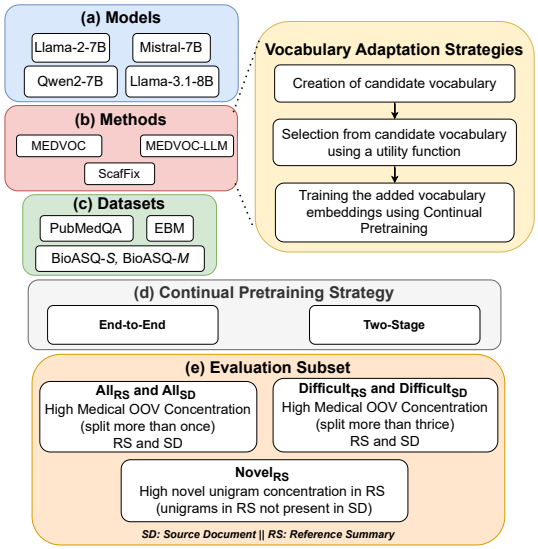

Figure 1: We present the benchmarking setup for our fine-grained evaluation of LLMs in high OOV and high novelty setting for medical text summarization. Our benchmarking setup: (a) tests four recent mainstream SoTA models; (b) on three vocabulary adaptation methods; (c) using three query-focused medical summarization dataset; (d) over two continual pretraining procedures; (e) evaluated in five fine-grained scenarios based on OOV (Out-Of-Vocabulary) and Novel unigram concentration in SD (Source Document) and RS (Reference Summary) respectively. So, in total we evaluated $4 \times 3 \times 4 \times 2 \times 5 = 480$ combinations.

*Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2025, pages 22989–23004

We note that LLMs during tokenization typically over-fragment (important) medical words, resulting in higher OOV concentration (proportion of words that are split into more than one token by an LLM tokenizer) compared to the generic open-domain text (e.g., news domain). In this work, we focus on the 'BASE' model variants of four open-source, open-domain LLMs called Llama-2 (Touvron et al., 2023), Mistal (Jiang et al., 2023), Qwen2 (Yang et al., 2024) and Llama-3.1 (Dubey et al., 2024). Llama-3.1 and Qwen2 has a much larger vocabulary size of 128K and 151K tokens, which is significantly larger than Llama-2 and Mistral (almost four times), which have a vocabulary size of around 32K. Therefore, it would be interesting to investigate whether vocabulary adaptation still helps to improve Llama-3.1 and Qwen2. We observe in Table 1, that there is an increase of 17.89%, 20.25%, 14.06%, and 13.08% in fragment score (Rust et al., 2021) when comparing tokenization of words from the open domain with the medical domain using Llama-2, Mistral, Qwen2, and Llama-3.1 tokenizers. This over-fragmentation affects the encoding stage as the semantic meaning is lost due to poor tokenization (Hofmann et al., 2022) and during generation, the model has to generate more subwords to generate a medical word. Examples of words that are over-fragmented is shown in Table 2. Additionally, summarization datasets in the medical domain being from a specialized domain contain words in reference summaries (like disease names) which do not occur directly in the source document but require domain-knowledge to be inferred from the source document (novel words).

Vocabulary expansion is a potential solution for such domain adaptation challenges that primarily arise due to vocabulary mismatch between the LLM vocabulary and the target expert-domain tasks. In this process, the LLM's vocabulary is expanded by incorporating tokens from the target domain that are not originally included in the LLM vocabulary. It has shown good performance for encoder-only models like BERT, RoBERTa on classification tasks (Hong et al., 2021; Hofmann et al., 2022; Lamproudis et al., 2022; Xu et al., 2023) and encoder-decoder models like BART, PEGASUS, and Transformers-Large on summarization and machine translation tasks (Xu et al., 2021; Nag et al., 2023; Balde et al., 2024b; Nag et al., 2024). Recent works are exploring vocabulary adaptation strategies for LLMs and proves useful for multilingual use-cases on models like Llama such as (Liu et al., 2023; Cui et al., 2024b,a; Liu et al., 2024; Gao et al., 2024; Nag et al., 2024; Tejaswi et al., 2024; Yamaguchi et al., 2024). However, these studies primarily focus on languages other than English, which are relatively underrepresented in the pretraining corpus of LLMs.

	CNN-DM	PAC
Llama-2		
Fragment Score	1.23	**1.45**
$\text{Split}_{>3}$	21%	**35%**
Mistral		
Fragment Score	1.17	**1.41**
$\text{Split}_{>3}$	21%	**34%**
Llama-3.1		
Fragment Score	1.07	**1.21**
$\text{Split}_{>3}$	12%	**25%**
Qwen2		
Fragment Score	1.28	**1.46**
$\text{Split}_{>3}$	17%	**28%**

Table 1: OOV concentration and fragment score observed for general domain dataset (CNN-DailyMail) and medical domain dataset (PAC) obtained using tokenizers of models: Llama-2 and Mistral (Vocabulary size: 32K), and Llama-3.1 and Qwen2 (Vocabulary size: 128K and 151K). Fragment score (Rust et al., 2021) is the average number of subwords a word is tokenized into; and $\text{Split}_{>3}$ is the fraction of words split more than thrice. We note that for all the model tokenizers, irrespective of the vocabulary size, medical domain words are over-fragmented leading to higher fragment score (highlighted in bold), compared to general domain.

We make the following contributions:

- We investigate vocabulary adaptation strategies in using LLMs on tasks from the expert (medical) domain. Instead of evaluating a multilingual setting, we address a vocabulary mismatch within the same language (English).
- We perform a benchmarking study on medical text summarization with various vocabulary adaptation strategies tailored for LLMs. We perform fine-grained evaluation of LLMs in high OOV and high novelty settings (Table 6).
- We conduct a human evaluation study with experts who also found that vocabulary adaptation produces more relevant, coherent, and faithful summaries (Figure 3).

2 Benchmarking Vocabulary Adaptation Strategies for LLMs

A vocabulary adaptation technique consists of three steps: (i) generating candidate vocabulary tokens

Medical Term	Llama-2	Llama-3.1
High OOV		
cardiomyopathy	'_card', 'iom', 'y', 'op', 'ath', 'y'	'card', 'i', 'omy', 'opathy'
antipyretics	'_ant', 'ip', 'y', 'ret', 'ics'	'ant', 'ipy', 'ret', 'ics'
High novelty		
corticosteroid	'_cort', 'ic', 'ost', 'ero', 'id'	'c', 'ortic', 'oster', 'oid'
antidepressant	'_ant', 'ide', 'press', 'ants'	'ant', 'idepress', 'ants'

Table 2: Medical terms from reference summary of PubMedQA dataset with high OOV concentration (*Difficult*$_{RS}$) and high novelty (*Novel*$_{RS}$).

from target downstream datasets, (ii) selecting important vocabulary tokens from candidate set using some utility function (e.g., fragment score (Hong et al., 2021), corpus entropy (Xu et al., 2021)) to form added vocabulary, and (iii) learning the embeddings of the added vocabulary and integration into LLM. In this work, we benchmark the effect of all the three steps, combining the effect of first two steps: constructing the vocabulary and adding candidate tokens from the target domain to finalize the vocabulary for integration into the LLM as one, and then checking effect of last step: on how to train the embeddings.

2.1 Vocabulary Adaptation Methods

We consider two types of datasets while constructing the vocabularies to be added to the PLM vocabulary: (i) PubMed Abstract Collection (PAC), a collection of around 300K PubMed abstracts used for intermediate fine-tuning by MEDVOC (Balde et al., 2024b), and (ii) Target Task (TGT Task), the target downstream task dataset for which the vocabulary is to be constructed. We now describe the different vocabulary adaptation methods used for our benchmarking study.

MEDVOC. MEDVOC (Balde et al., 2024b) is a SoTA vocabulary adaptation strategy for adapting PLMs like BART and PEGASUS, on medical summarization tasks. First, candidate vocabularies are constructed on the medical OOV words from a domain-specific corpus (PAC) – V_{PAC}, and a downstream task dataset –V_{TGT}. Then, an optimal vocabulary (V_{MEDVOC}), that lies at the intersection of V_{PAC} and V_{TGT} is chosen via a hyperparameter search. The utility function for the hyperparameter search is fragment score (Rust et al., 2021), defined as the average number of subwords a word from a corpus $\mathcal{C}$ is tokenized into by a tokenizer using vocabulary $\mathcal{V}$. The vocabulary configuration within the neighborhood of the optimal vocabulary is finally chosen to avoid overfitting on large vocabulary sizes.

MEDVOC-LLM. This is a variant of MEDVOC adapted for LLM vocabularies and tokenizers. We identify two key issues in MEDVOC: (i) many of the vocabulary terms added directly from V_{PAC}, did not occur even once in the reference summaries of the train set of downstream target task – their addition did not contribute during generation, and (ii) certain added vocabulary terms were a mixture of numerals and punctuations (e.g., *-9,*) – which is not consistent with the tokenization scheme for LLMs considered in this study[1] (Llama-2 and Llama-3.1). Thus, in this approach, we clean the vocabularies generated by MEDVOC by removing the terms from both categories.

Overhead in Previous Vocabulary Adaptation Strategy. Consider the word 'cholesterol', which is not present in the Llama-3.1 model's vocabulary and is tokenized by the Llama-3.1 tokenizer as *[cho, le, sterol]*. Since a merge rule operates on pairs of tokens, we need to iteratively add pairs from left to right, as shown in Table 3.

Token	Merge Rule
'chole'	[ch, ole]
'cholesterol'	[chole, sterol]

Table 3: Illustration of iterative addition of tokens to add a target token '*cholesterol*' in the vocabulary.

Thus, in order to add 'cholesterol' to the vocabulary, we need to add one extra token [chole]. Specifically, for Llama-3.1, we observe an overall addition of 20% of extra such tokens, which are seldom used once the complete word is added into the vocabulary. This can lead to reduced performance in downstream tasks. To that end, we present ScafFix to address this issue with existing vocabulary adaptation strategies.

ScafFix. Unlike the vocabulary adaptation strategies described till now, ScafFix constructs the candidate set for added vocabulary by directly considering the medical words and ignores the tok-

[1]Llama tokenizers explicitly set apart digits as individual tokens (Touvron et al., 2023)

enization step for forming candidate subwords. We directly select x tokens (where x represents the quota, set to 500 in this case, in steps of 50) from the candidate vocabulary tokens, ranking them by their frequency in descending order. We then follow the MEDVOC-based hyperparameter search optimizing fragment score to obtain the optimal vocabulary to be added. To offset the absence of such derivative tokens, we use AdaptBPE tokenization scheme (Balde et al., 2024a) instead of the standard model tokenizers. Instead of directly utilizing merge rules, AdaptBPE first checks whether a part of the input token (using longest-first match) directly exists in the added vocabulary, preserves it from splitting, and then runs the merge-based byte-pair encoding scheme iteratively on the rest of the input. Here, we avoid including the scaffolding tokens (Cognetta et al., 2024; Bauwens and Delobelle, 2024; Chizhov et al., 2024; Lian et al., 2025) during the vocabulary addition phase; these are derivative tokens that remain under-trained once the whole word is added to the vocabulary.

2.2 Learning the Added Vocabulary Embeddings using Continued Pretraining

The embedding of the newly added vocabulary is initialized as the average of the embeddings of the existing subwords in the vocabulary (Yamaguchi et al., 2024). We explore two continual pretraining (Gururangan et al., 2020; Tejaswi et al., 2024; Yamaguchi et al., 2024) strategies to train these embeddings on target domain text (20K random documents from the PubMed Abstract Collection, in our case). Continual pretraining uses the same training objective of 'Next Token Prediction' as the autoregressive language modeling objective, and optimizes the standard negative log-likelihood loss. We use the popular parameter-efficient fine-tuning technique known as 'Low-Rank Adaptors' (LoRA) (Mangrulkar et al., 2022) because end-to-end training of LLMs is computationally infeasible. We explore two continual pretraining strategies:

- **End-To-End**: The model is trained in an end-to-end manner by freezing all the base model layers except the input and output embedding layers and training LoRA adapters.
- **Two-Stage**: First, the entire model is frozen along with the LoRA layers except for the input and output embedding layers for a short duration, then unfreeze the LoRA adapters and train the LoRA adapters along with the embedding layers (Cui et al., 2024b; Yamaguchi et al., 2024). The second approach leads to more stable training and avoids overfitting to the initial LLM embedding space.

Category	Description
Categories of Words	
Difficult-OOV	Medical words that are split more than thrice by model tokenizers
Novel	Words in the summary that are not present in the source document
All-OOV	Medical words that are split more than once by model tokenizers
Evaluation Setting (Top ten percentile)	
$\text{Difficult}_{\text{RS}}$	High Difficult-OOV concentration in the reference summaries
$\text{Difficult}_{\text{SD}}$	High Difficult-OOV concentration in the source document
Novel_{RS}	High Novel concentration in the reference summaries
All_{SD}	High All-OOV concentration in the source document
All_{RS}	High All-OOV concentration in the reference summaries

Table 4: Challenging fine-grained evaluation scenarios considered in this study. We focus on the subset that contains high OOV concentration and novelty

3 Experimental Setup

We use in-context learning (ICL) (Brown et al., 2020; Van Veen et al., 2024) along with greedy decoding for generating summaries from LLMs. We follow the approach similar to ClinSumm (Van Veen et al., 2024), where examples are sampled from the train set using the similarity computed using PubMedBERT[2] with the given test data point. The template for prompting is shown in Appendix A (Table 9). We also describe the fine-grained evaluation setup used for this benchmarking study in Table 4.

Next, we describe the benchmark medical text summarization datasets used, followed by details on evaluation metrics, baseline models, and training details.

3.1 Datasets

In this work, we focus on three query-focused medical summarization datasets such as BioASQ (Tsatsaronis et al., 2015), EBM (Mollá and Santiago-Martinez, 2011), and PubMedQA (Jin et al., 2019).

[2] https://huggingface.co/pritamdeka/PubMedBERT-mnli-snli-scinli-scitail-mednli-stsb

A data point consists of a Query (Q) and a context (PubMed Abstract) as the input to the model. The gold-standard reference summary (RS) summarizes the abstract based on the query. We present the overall dataset characteristics in Table 5.

PubMedQA. PubMedQA is a question-answering dataset with 1000 human-annotated data points. Here, input is a query appended to a PubMed abstract, which forms the source document. We consider the 'long answer' as the reference summaries.

EBM. In this single-document summarization task, the input comprises a query paired with a PubMed abstract. The task is to generate a concise summary that addresses the query, using the provided abstract as its context.

BioASQ. We use the Phase-B query-focused summarization task of BioASQ-9B. The input includes a question along with relevant PubMed abstracts. For the summarization task, we use the ideal answer as the reference summary. Here, we explore two variants of a source document — (i) Snippets (BioASQ-$\mathcal{S}$): the question followed by the list of relevant snippets from a collection of PubMed Abstracts, in line with the MEDVOC paper (Balde et al., 2024b), and (ii) BioASQ-Main Abstract (BioASQ-$\mathcal{M}$): the question followed by the complete Pubmed abstracts.

3.2 Evaluation Metrics

We evaluate the model-generated summaries using Rouge-L (R-L) to measure informativeness and coherence, and Concept-Score (CSr) to measure faithfulness (Zhang et al., 2023). Concept-Score measures the overlap of UMLS medical concepts (computed using QuickUMLS tool (Soldaini and Goharian, 2016)) between the generated and reference summaries. We use Rouge-L as the primary comparison metric, in line with prior studies (Fabbri et al., 2021; Yuan et al., 2022; Balde et al., 2024a,b).

3.3 Baseline Models

The following baseline models do not update the LLM vocabulary:

- **BASE.** 'BASE' corresponds to the base variant of original LLM without any vocabulary adaptation and continual pretraining. We consider 7B variant of Llama-2 (Model id: meta-llama/Llama-2-7b-hf), Mistral (Model id: mistralai/Mistral-7B-v0.1), Qwen2 (Model id: Qwen/Qwen2-7B), and 8B variant of Llama-3.1 (Model id: meta-llama/Llama-3.1-8B) as our BASE models.
- **Continual Pretraining (CPT-Only).** CPT-Only corresponds to the original BASE LLM that has undergone the continual pretraining (CPT) but without any vocabulary adaptation. It serves as a strong baseline when comparing against vocabulary-adapted models.

3.4 Implementation Details

We provide basic details of implementation such as details on LoRA, size of pretraining corpus, and training hyperparameters.

Continual Pretraining using LoRA. We use one A100 40 GB GPU to carry out the pretraining. We use LoRA (Hu et al., 2022; Mangrulkar et al., 2022) to carry out the pretraining in this resource-constrained setting. The LoRA adapters are applied to all the linear modules in the model. These include $\{k\_proj,\ q\_proj,\ v\_proj,$ and $o\_proj\}$ modules from self attention layers along with $\{gate\_proj,\ up\_proj,$ and $down\_proj\}$ modules from MLP layers. We use a consistent LoRA configuration of a rank value of 32 and an alpha value of 64 for fair comparison.

Size of Pretraining Corpus. As continual pretraining with all the 312K documents from the PubMed Abstracts Collection (PAC) is computationally infeasible for LLMs, we perform a hyperparameter search to identify an optimal dataset size. We experiment with various dataset sizes of 10K, 20K, 50K, and 100K, and observe that the performance observed for 20K which only takes 6 hours of pretraining was comparable with the of 100K setting, which took approximately 40 hrs of pretraining. Therefore, we *continually pretrain over the BASE model with randomly selected 20K documents from PAC*.

Training Hyperparameters. We use a global batch size of 32 (on device: 8 with gradient accumulation: 4), and a learning rate of $1e-4$. In End-to-End pretraining procedure, LoRA layers of the model were trained end-to-end for 5 epochs. In case of Two-Stage pretraining procedure, the embedding layers training phase and keeping LoRA layers frozen, was carried out for 2 epochs on 10K PAC samples. Then both the LoRA and embedding layers were trained end-to-end on 20K PAC

Dataset	Test Set Size	Token Count of Reference Summaries		OOV Concentration Split more than once (in %)				OOV Concentration Split more than thrice (in %)				Unigram Novelty (in %)
		Llama-2	Llama-3.1	Llama-2		Llama-3.1		Llama-2		Llama-3.1		
				SD	RS	SD	RS	SD	RS	SD	RS	
PubMedQA	500	63	25	36.67	38.00	43.68	45.65	4.91	4.65	2.61	2.42	41.32
EBM	424	112	91	38.97	40.90	45.60	46.23	6.65	7.92	3.90	5.17	47.15
BioASQ-$\mathcal{M}$	963	85	69	46.20	50.64	52.03	56.61	9.12	11.04	5.55	7.09	42.58
BioASQ-$\mathcal{S}$	496	73	58	47.12	50.00	52.00	57.15	8.70	9.10	4.76	4.55	4.11

Table 5: Medical text summarization datasets used for evaluation. We have three key observations: (i) BioASQ has the highest OOV concentration; (ii) EBM has highest novelty concentration; and (iii) EBM has the longest length reference summaries.

samples for 3 epochs. In both the pretraining procedures, the base layers of the models were kept frozen throughout the training process.

4 Experimental Results

Table 6 shows the performance comparison of the different vocabulary adaptation strategies on Llama-2 and Llama-3.1 models. We report the best performance among the two continual pretraining variants of 'End-to-End' and 'Two-Stage' as previously described in Section 2.2. The performance values of the individual settings is added to the Appendix B. Since Mistral vocabulary size is same as Llama-2 (32K), and Qwen2 vocabulary size is similar as Llama-3.1 (151K and 128K), we add the Mistral (Table 13) and Qwen2 (Table 14) results to the Appendix B. We summarize the key results for Qwen2 and Mistral model in RQ7. We observe that vocabulary adaptation leads to performance improvement in terms of Rouge-L for Llama-2 and Llama-3.1 in seven out of eight settings (except for Llama-2 on EBM dataset). In terms of Concept-Score which is a proxy measure for faithfulness (Zhang et al., 2023), Figure 2 shows that at least one vocabulary adaptation performs the best in five out of six settings, except for Llama-3.1 on PubMedQA dataset.

4.1 Discussion of Results for Vocabulary Adaptation with LLMs

We focus on the fine-grained analysis of LLMs in high OOV and high novelty data points. The Rouge-LCS values observed for each such category are reported in Table 6. In total, we benchmark an LLM on 12 settings: 8 OOV-related settings and 4 novelty-related settings. Here we report the best performing training method in $\text{Difficult}_{\text{SD}}$, $\text{Difficult}_{\text{RS}}$ and Novel_{RS} setting. The complete results can be found in Table 11 (for Llama-2) and Table 12 (for Llama-3.1). The Concept score values are shown in Figure 2. We find that at least one vocabulary adaptation strategy improves over BASE in a total of five out of six settings (LLM-Dataset pairs). The overall average improvement for 18.75% for Llama-2 and 14.82% for Llama-3.1 models across datasets. We present two representative examples from EBM and PubMedQA datasets using Llama-2 models in Appendix D.

Now, we explore research questions to evaluate LLMs in difficult settings and scenarios where vocabulary adaptation strategies does not help.

RQ1: Vocabulary adaptation outperform BASE model on full test data ($\text{Test}_{\text{Full}}$). We note that at least one of the vocabulary adaptation strategies (best of MEDVOC-LLM and ScafFix) improves over BASE in 5 out of 8 settings. The average performance improvement is 3.68% for Llama-2 and 4.57% for Llama-3.1. The methods where we see improvements are: MEDVOC-LLM (2/5) and ScafFix (3/5) settings. Thus on the entire test set, **ScafFix**, is the best-performing vocabulary adaptation method. It even outperforms CPT-Only in six out of eight settings by a margin of 2.97%.

RQ2: CPT-Only improves over BASE in high novelty and OOV concentration. We find that CPT-Only (Continual Pretraining without any vocabulary adaptation) improves over BASE in 4 out of 8 (Llama-2: 2; Llama-3.1: 2) high novelty settings. CPT-Only improves over BASE in 11 out of 16 (Llama-2: 6; Llama-3.1: 5) higher OOV settings. However, we observe that at least one vocabulary adaptation strategy improves over CPT-Only in high OOV (13 out of 16 settings) and high novelty (in all 8 setting); thus necessitating the need for vocabulary adaptation.

RQ3: Vocabulary adaptation helps in high OOV concentrations in reference summaries and source documents. We observe that vocabulary adaptation (best of MEDVOC-LLM and ScafFix) outperforms BASE in 14 out of 16 set-

Model	Test$_{Full}$	Difficult$_{RS}$	Novel$_{RS}$	Difficult$_{SD}$	Overall	Test$_{Full}$	Difficult$_{RS}$	Novel$_{RS}$	Difficult$_{SD}$	Overall
	Llama-2					Llama-3.1				
PubMedQA										
BASE	26.33	24.40	19.78	23.53	23.51	**28.10**	26.87	18.87	24.46	24.58
CPT-only	27.12	28.00	22.21	25.97	25.83	26.62	**27.08**	21.07	24.52	24.82
MEDVOC	26.50	22.88	20.14	24.35	23.47	27.86	26.23	21.87	24.68	25.16
MEDVOC-LLM	26.90	26.67	20.00	24.49	24.52	27.69	26.67	21.74	**26.51**	**25.65**
ScafFix	**27.61**	**28.57**	**22.22**	**26.32**	**26.18**	27.67	25.00	**22.22**	23.85	24.69
EBM										
BASE	18.56	**16.00**	11.20	17.33	15.77	20.04	14.54	11.37	15.79	15.44
CPT-only	19.13	15.87	11.77	**19.05**	**16.46**	20.13	17.15	13.33	16.98	16.90
MEDVOC	18.60	15.38	12.50	18.65	16.28	20.31	16.03	12.17	17.93	16.61
MEDVOC-LLM	**19.27**	14.71	**12.90**	17.54	16.11	20.75	16.40	13.04	**18.39**	**17.15**
ScafFix	18.65	14.71	11.77	17.40	15.63	**20.79**	**17.17**	**13.80**	16.18	16.99
BioASQ-$\mathcal{S}$										
BASE	32.12	26.12	20.53	35.72	28.62	35.25	29.60	18.18	27.03	27.52
CPT-only	**33.30**	29.22	20.00	30.33	28.21	36.01	24.70	17.14	29.41	26.82
MEDVOC	32.26	28.17	18.18	27.40	26.50	37.01	29.79	18.18	32.43	29.35
MEDVOC-LLM	32.40	**29.41**	20.29	30.33	28.11	**37.15**	**32.89**	20.29	**37.84**	**32.04**
ScafFix	32.88	29.29	**22.22**	**36.44**	**30.21**	36.70	32.26	**20.90**	34.15	31.00
BioASQ-$\mathcal{M}$										
BASE	**28.50**	27.27	**21.53**	28.00	26.33	**29.28**	26.67	**21.51**	28.57	26.51
CPT-only	27.22	27.91	20.31	28.57	26.00	27.56	26.09	19.84	28.57	25.52
MEDVOC	24.19	23.26	16.53	26.32	22.73	27.71	27.27	20.46	28.57	26.01
MEDVOC-LLM	24.50	25.00	18.90	28.00	24.10	27.45	27.27	19.05	29.17	25.74
ScafFix	26.16	**29.79**	21.17	**30.00**	**26.78**	28.91	**27.59**	21.23	**30.00**	**26.94**

Table 6: Fine-grained performance evaluation of vocabulary adaptation strategies and baseline modes in terms of Rouge-L (**R-L**). The values represent the best among the two continual pretraining strategies of 'End-to-End' and 'Two-Stage'. The 'Overall' column represents the average value of Test$_{Full}$, Difficult$_{RS}$, Novel$_{RS}$ and Difficult$_{SD}$. Vocabulary adaptation (best of MEDVOC-LLM and ScafFix) improves over BASE in 7 out of 8 overall settings.

tings. For these settings, the average performance improvement over BASE is 8.74% and 14.64% for Llama-2 and Llama-3.1 respectively. Thus, **MEDVOC-LLM** is the best vocabulary adaptation strategy in scenarios of high OOV concentration in reference summaries and source documents.

RQ4: Vocabulary adaptation helps in high novelty settings. We observe that vocabulary adaptation improves over BASE in 6 out of 8 high novelty settings. The average performance improvement (wherever observed) is 11.92% and 18.03% for Llama-2 and Llama-3.1 when compared to BASE. Thus, **ScafFix** is the best vocabulary adaptation strategy in high novelty settings as it outperformed in five out of eight settings.

Word	BASE tokenization	ScafFix tokenization
microbiologically	['Ġmicrobi', 'ologically']	['*Ġmicro*', *'biological'*, *'ly'*]
inhibitory	['Ġinhib', 'itory']	['*Ġ*', *'inhibitor'*, *'y'*]
chronically	['Ġchron', 'ically']	['*Ġ*', *'chronic'*, *'ally'*]
antibacterial	['Ġantib', 'acterial']	['*Ġanti'*, *'bacteria'*, *'l'*]

Table 7: Samples of words from PubMedQA corpus. For each word, we observe that ScafFix tokenization preserves the morphological boundary while tokenization, unlike BASE tokenization where subwords cross morphological boundaries (Bauwens and Delobelle, 2024).

RQ5: Comparison of added vocabulary sizes among vocabulary adaptation strategies. The details of the added vocabulary and fragment score for the different vocabulary adaptation strategies are provided in Table 8 (Appendix A). We observe that ScafFix reduces fragment score by 29.19% over BASE by adding the minimum amount of added vocabulary terms as compared to the MEDVOC and MEDVOC-LLM. The lowest vocabulary size due to removal of scaffolding tokens leads to a significant drop in the number of under-trained tokens in the LLM vocabulary. This makes the training phase less noisy and yields superior performance during inference.

RQ6: Vocabulary adaptation does not help much in case of extractive summaries, low OOV concentration and low novelty. We note that all the vocabulary adaptation techniques struggle to outperform both BASE and CPT-Only in PubMedQA (Llama-3.1), BioASQ-$\mathcal{S}$ (Llama-2), and BioASQ-$\mathcal{M}$ (Llama-2 and Llama-3.1). When compared in the entire test set; there is a performance drop of 1.39% across these two settings. This may be because PubMedQa and BioASQ-$\mathcal{S}$ have low novelty. Specifically, for BioASQ-$\mathcal{S}$, the unigram novelty is just 4.11% (see Table 5), therefore, the high novelty threshold for top-ten percentile was 35%. Thus, we take the average of EBM and PubMedQA high novelty values of 60%. This results

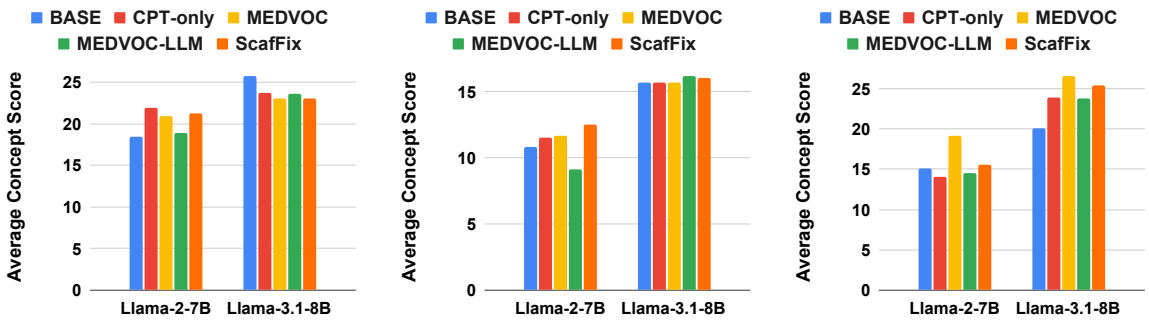

Figure 2: Concept Score (CSr) observed for **(a) PubMedQA**; **(b) EBM**; and **(c) BioASQ-$\mathcal{M}$** averaged over $\text{Difficult}_{\text{SD}}$, $\text{Difficult}_{\text{RS}}$, and Novel_{RS} setting. At least one vocabulary adaptation strategy imprves over BASE in a total of 5 out of 6 comparisons. The overall average improvement for 18.75% for Llama-2 and 14.82% for Llama-3.1 models across datasets.

in vocabulary adaptation outperforming BASE in high novelty setting of BioASQ-$\mathcal{S}$.

In case of BioASQ-$\mathcal{M}$, we note that although it is a dataset with higher unigram novelty, the summaries where ScafFix fails have a higher Rouge-L overlap with the Source Document as compared to the summaries where ScafFix does better than BASE, thus making the inference easier irrespective of OOV and Novelty concentration. We provide a detailed error analysis observed in the performance gap in Appendix B (Table 10)

RQ7: Proposed vocabulary adaptation methods generalizes to other LLMs such as Qwen2 and Mistral. In case of Qwen2 (Table 14), we observe that at least one vocabulary adaptation strategies improve over BASE and CPT-Only in 7 out of 12 comparisons. In case of Mistral (Table 13), we observe that at least one of the vocabulary adaptation strategies improved over BASE and CPT-Only in one out of 12 comparisons. In case of both the models (details in Appendix B), we observe a similar trend with vocabulary adaptation not helping in case of low OOV and novelty concentration (specifically in BioASQ-$\mathcal{S}$) as we observed in RQ6. These findings highlight that our results are potentially generalizable to multiple LLMs.

4.2 Human Evaluation

We randomly select 30 test data points uniformly across the three datasets and two models that have higher expert OOV concentration. We use the Prolific platform to recruit medical experts for annotating summary pairs of ScafFix model and BASE across three aspects (Fabbri et al., 2021; Zhang et al., 2023; Balde et al., 2024b) namely *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 C for more details), and each summary par was evaluated by three annotators. Figure 3 shows the human evaluation results where the ScafFix method generates more faithful summaries (93.34% versus 83.34% of summaries are faithful), and more relevant summaries, where 93.34% of data points get a score ≥ 4 in Likert scale, as compared to 70% by BASE.

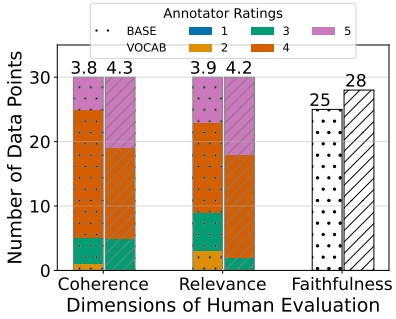

Figure 3: Human evaluation scores in high OOV concentration on 'End-to-End' continual pretraining strategy. Vocab corresponds to **ScafFix** model which produces more relevant, coherent, and faithful summaries during human evaluation with medical experts.

5 Related Works

Vocabulary Expansion in LLMs. Recent research has focused on enhancing large language models (LLMs) through domain-specific vocabulary optimization. Liu et al. (2024) introduces VEGAD, an adaptive method for selecting an opti-

mal subset of domain vocabulary, which enhances performance on both specialized and general tasks, validated on Chinese datasets. Similarly, Gao et al. (2024) introduced VE-KD, a method combining vocabulary expansion and knowledge distillation to train efficient domain-specific language models for BioBERT and PubMedBERT on biomedical tasks. These approaches align with broader efforts, such as, Chatlaw (Cui et al., 2024a), an AI legal assistant, that employed a knowledge graph-enhanced mixture-of-experts model to address legal domain challenges, and (Cui et al., 2024b), who developed efficient text encoding strategies for Chinese LLMs. Similar efforts have been put in other languages (Nag et al., 2024; Tejaswi et al., 2024; Yamaguchi et al., 2024). However, none of the works benchmark the effect in the fine-grained manner that we aimed to show in this study.

Over-fragmentation in Medical domain. Over-fragmentation (spliting of domain words in more than one subword) in tokenization is a significant challenge in adapting large language models (LLMs) to the medical domain due to specialized terminology. Si et al. (2019) highlighted the limitations of traditional tokenization on the task of clinical concept extraction where they show models like BERT poorly toknize medical named entities. Nguyen et al. (2019) emphasized the impact of lexical segmentation on transformer-based models and suggested that not only using a domain-specific vocabulary helps, but also a continual pretraining phase helps. Yuan et al. (2022) introduced BioBART and Labrak et al. (2024) introduced BioMistral, biomedical-specific generative model designed to address domain-specific tokenization challenges that outperformed various SoTA methods on different kinds of task like summarization and question-answering. Liu et al. (2023) proposed task-adaptive tokenization to enhance long-form text generation by dynamically adjusting tokenization for domain-specific semantics by going beyond word boundaries during tokenization. Our work builds on these advancements by benchmarking the effect of refining tokenization on recent Llama-2, Llama-3.1, Mistral, and Qwen2 models.

6 Conclusion

This work is a first step towards understanding how vocabulary adaptation strategies effect the performance of general-purpose LLMs in medical text summarization. We first show that general-purpose LLMs fail in certain challenging generation scenarios where reference summaries have high OOV concentration and high novelty. We then benchmark the performance of three vocabulary adaptation strategies on four models: Llama-2 7B, Mistral 7B, Qwen2 7B, and Llama-3.1 8B model; over three biomedical summarization datasets and two continual pretraining strategies. Llama-3.1 and Qwen2 even with a vocabulary size of 128K and 151K tokens, still faces the over-fragmentation issue with medical words, and vocabulary adaptation is shown to help improve the LLM summarization performance. Medical experts find that vocabulary adaptation improves the relevance, coherence, and faithfulness of medical summaries. We make the codebase available for reproducibility purposes[3].

7 Limitations

We identify two main limitations of this work. Firstly, because of hardware constraints we could not explore much larger variants (like 13B and 70B) of the models considered in the study. Second, we note that all the results provided in this paper are generated by applying LoRA during the pretraining phase. This might make a difference when replicating the results with full-scale fine-tuning without any parameter efficient fine-tuning strategies.

8 Ethics Statement and Broader Impact

It is a well known fact that LLMs are often prone to hallucinations, producing outputs that could not be verified as-is from the source context without any domain-specific knowledge. While the proposed methods here do improve faithfulness, the quality of the summaries generated is still not ready for deployment to the public without evaluating the safety perspective of such responses. We believe more research is needed to align the outputs of these LLMs for such high-stake domains like healthcare where a single misinformation (or disinformation) could lead to drastic consequences.

Acknowledgments

We thank the Ministry of Education, Govt of India, for supporting Gunjan Balde with Prime Minister Research Fellowship during his Ph.D. tenure. This research was partially funded by a Google Academic Research Award.

[3] https://github.com/gb-kgp/LLM-MedicalSummarization-Benchmark

References

Gunjan Balde, Soumyadeep Roy, et al. 2024a. Adaptive BPE tokenization for enhanced vocabulary adaptation in finetuning pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14724–14733, Miami, Florida, USA. Association for Computational Linguistics.

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

Thomas Bauwens and Pieter Delobelle. 2024. BPE-knockout: Pruning pre-existing BPE tokenisers with backwards-compatible morphological semi-supervision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5810–5832, Mexico City, Mexico. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Pavel Chizhov, Catherine Arnett, et al. 2024. BPE gets picky: Efficient vocabulary refinement during tokenizer training. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16587–16604, Miami, Florida, USA. Association for Computational Linguistics.

Marco Cognetta, Tatsuya Hiraoka, et al. 2024. An analysis of BPE vocabulary trimming in neural machine translation. In *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, pages 48–50, Mexico City, Mexico. Association for Computational Linguistics.

Jiaxi Cui, Munan Ning, et al. 2024a. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *Preprint*, arXiv:2306.16092.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024b. Efficient and effective text encoding for chinese llama and alpaca. *Preprint*, arXiv:2304.08177.

Tim Dettmers, Artidoro Pagnoni, et al. 2024. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Pengju Gao, Tomohiro Yamasaki, and Kazunori Imoto. 2024. Ve-kd: Vocabulary-expansion knowledge-distillation for training smaller domain-specific language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15046–15059.

Suchin Gururangan, Ana Marasović, et al. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Valentin Hofmann, Hinrich Schuetze, and Janet Pierrehumbert. 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, Dublin, Ireland. Association for Computational Linguistics.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Edward J Hu, yelong shen, et al. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

Qiao Jin, Bhuwan Dhingra, et al. 2019. Pubmedqa: A dataset for biomedical research question answering. 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 2567–2577.

Yanis Labrak, Adrien Bazoge, et al. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Andrew Lampinen, Ishita Dasgupta, et al. 2022. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anastasios Lamproudis, Aron Henriksson, et al. 2022. Vocabulary modifications for domain-adaptive pretraining of clinical language models. In *Proceedings of the 15th International Joint Conference on*

Biomedical Engineering Systems and Technologies (BIOSTEC 2022) - HEALTHINF, pages 180–188. INSTICC.

Haoran Lian, Yizhe Xiong, et al. 2025. Scaffold-bpe: Enhancing byte pair encoding for large language models with simple and effective scaffold token removal. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24539–24548.

Chengyuan Liu, Shihang Wang, et al. 2024. Gold panning in vocabulary: An adaptive method for vocabulary expansion of domain-specific LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7442–7459, Miami, Florida, USA. Association for Computational Linguistics.

Siyang Liu, Naihao Deng, et al. 2023. Task-adaptive tokenization: Enhancing long-form text generation efficacy in mental health and beyond. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15264–15281, Singapore. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, et al. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

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

Arijit Nag, Animesh Mukherjee, et al. 2024. Cost-performance optimization for processing low-resource language tasks using commercial LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15681–15701, Miami, Florida, USA. Association for Computational Linguistics.

Arijit Nag, Bidisha Samanta, et al. 2023. Entropy-guided vocabulary augmentation of multilingual language models for low-resource tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8619–8629, Toronto, Canada. Association for Computational Linguistics.

Vincent Nguyen, Sarvnaz Karimi, and Zhenchang Xing. 2019. Investigating the effect of lexical segmentation in transformer-based models on medical datasets. In *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, pages 165–171.

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, Online. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Yuqi Si, Jingqi Wang, et al. 2019. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26(11):1297–1304.

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

Atula Tejaswi, Nilesh Gupta, and Eunsol Choi. 2024. Exploring design choices for building language-specific LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10485–10500, Miami, Florida, USA. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv 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*, 16:1–28.

Dave Van Veen, Cara Van Uden, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

Benfeng Xu, Chunxu Zhao, et al. 2023. Retrieval-augmented domain adaptation of language models. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 54–64.

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, Online. Association for Computational Linguistics.

Atuki Yamaguchi, Aline Villavicencio, and Nikolaos Aletras. 2024. How can we effectively expand the vocabulary of llms with 0.01 gb of target language text? *arXiv preprint arXiv:2406.11477*.

An Yang, Baosong Yang, et al. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Hongyi Yuan, Zheng Yuan, et al. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Nan Zhang, Yusen Zhang, et al. 2023. Famesumm: Investigating and improving faithfulness of medical summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10915–10931.

Weijia Zhang, Jia-Hong Huang, et al. 2025. Beyond relevant documents: A knowledge-intensive approach for query-focused summarization using large language models. In *International Conference on Pattern Recognition*, pages 89–104. Springer.

A Experimental Setup

Vocabulary Sizes. We report the vocabulary sizes along with the resultant fragment score in Table 8.

Model	Llama-2		Mistral		Llama-3.1		Qwen2	
	|V|	FragSr	|V|	FragSr	|V|	FragSr	|V|	FragSr
PubMedQA								
CPT-Only	32000	2.53	32000	2.58	128256	2.29	151646	2.29
MEDVOC	38245	1.48	37642	1.55	129078	1.21	152441	1.21
MEDVOC-LLM	34572	1.66	34219	1.73	128808	1.25	152131	1.25
ScafFix	32200	2.14	32266	2.20	128456	1.31	152000	1.28
EBM								
CPT-Only	32000	2.65	32000	2.75	128256	2.33	151646	2.33
MEDVOC	42836	1.48	43908	1.61	131231	1.12	154092	1.13
MEDVOC-LLM	34572	1.66	36252	1.80	130253	1.22	153468	1.23
ScafFix	32150	1.83	32060	2.54	128456	1.37	152062	1.35
BioASQ								
CPT-Only	32000	2.98	32000	2.69	128256	2.44	151646	2.65
MEDVOC	43194	1.93	46502	1.44	133099	1.24	155518	1.26
MEDVOC-LLM	37399	2.18	37506	1.65	130966	1.30	154612	1.30
ScafFix	32300	2.43	32193	2.39	128506	1.56	152098	1.44

Table 8: The final vocabulary sizes (|V|) along with the resultant fragment score (**FragSr**) observed on Medical OOV words. MEDVOC, MEDVOC-LLM, and ScafFix are the best vocabulary adaptation scheme which has the least vocabulary sizes and decent fragment score.

Prompts used. We use the prompt template inspired from ClinSumm (Van Veen et al., 2024) and Zhang et al. (2025) as shown in Table 9.

B Results

Comparing Continual Pretraining Strategies for Llama-2 and Llama-3.1. In Table 11, we report the performance we observe using two different continual pretraining strategies on Llama-2 and in Table 12 we show performance for Llama-3.1 model. In case of Llama-2 the best pretraining strategy was Two-Stage and and for Llama-3.1 was End-To-End.

Mistral and Qwen results. We present the Mistral results in Table 13 and Qwen results in Table 14.

In case of Qwen2, we observe that at least one vocabulary adaptation strategies improve over BASE and CPT-Only in 7 out of 12 comparisons considered in the main text ($Difficult_{SD}$, $Difficult_{RS}$, $Novel_{RS}$). Here, we observe a similar trend with vocabulary adaptation not helping in case of low OOV and novelty concentration (specifically in BioASQ-$\mathcal{S}$) as we observed in main text (RQ6). The best performing continual pretraining strategy was End-to-End.

In case of Mistral, we observe that none of the vocabulary adaptation strategies improved over BASE and CPT-Only in one out of 12 comparisons considered in the main text ($Difficult_{SD}$, $Difficult_{RS}$, $Novel_{RS}$). Here, we observe a similar trend with vocabulary adaptation not helping in case of low OOV and novelty concentration (specifically in BioASQ-$\mathcal{S}$) as we observed in main text (RQ6). The best performing continual pretraining strategy was End-to-End for all the datasets except EBM.

Error Analysis for Performance Gap in BioASQ-$\mathcal{M}$. We split the test set into two subsets: (i) Subset-1: instances where BASE has better Rouge-LCS than ScafFix (512 data points), and (ii) Subset-2: instances where ScafFix has higher Rouge-LCS than BASE (444 data points). We analyzed characteristics like fraction of Medical OOV words in Ref-

Prompt
You are a medical expert. You are given a query and query-relevant information as inputs. Your task is to summarize this information. The summary should be concise, include only non-redundant, query-relevant evidence, and be approximately 100 words long. Use the provided examples to guide word choice.
–n icl examples concatenated using '##'–
Query {i}: {Train-Query}
Document {i}: {Train-Source Document}
Summary {i} : {Train-Summary}
##
–Test Example–
Query: {Test-Query}
Document: {Test-Source Document}
Summary:

Table 9: Prompt Template used to prompt the language models for the task of query focused summarization.

Metric	Subset-1	Subset-2
Llama-2		
Difficult-RS Concentration	9.78%	10.77%
Novel-RS Concentration	40.65%	41.92%
Rouge-LCS overlap between source and reference	38.17	35.65
Llama-3.1		
Difficult-RS Concentration	7.47%	8.81%
Novel-RS Concentration	39.46%	42.78%
Rouge-LCS overlap between source and reference	39.52	34.27

Table 10: Difference of characteristics between instances where BASE has better Rouge-LCS than ScafFix (Subset-1) and instances where ScafFix has higher Rouge-LCS than BASE (Subset-2). The instances where BASE has better Rouge-LCS than ScafFix (Subset-1) have lesser values of Difficult-RS Concentration as well as Novel-RS Concentration, but higher values of Rouge-LCS overlap compared to Subset-2.

	End-To-End						Two-Stage					
	$\text{Test}_{\text{Full}}$	$\text{Difficult}_{\text{SD}}$	$\text{Difficult}_{\text{RS}}$	Novel_{RS}	All_{SD}	All_{RS}	$\text{Test}_{\text{Full}}$	$\text{Difficult}_{\text{SD}}$	$\text{Difficult}_{\text{RS}}$	Novel_{RS}	All_{SD}	All_{RS}
PubMedQA												
BASE	26.33	23.53	24.40	19.78	25.46	27.45	26.33	23.53	24.40	19.78	22.64	27.45
CPT-only	27.12	25.97	28.00	22.21	**27.86**	27.59	27.11	25.00	26.92	21.24	**24.07**	26.67
MEDVOC	26.50	24.35	22.88	20.14	24.59	27.70	26.39	24.00	24.82	20.88	22.22	26.23
MEDVOC-LLM	26.90	24.49	26.67	20.00	25.27	27.12	**27.30**	**25.54**	**27.45**	20.84	21.43	28.89
ScafFix	**27.61**	**26.32**	**28.57**	**22.22**	25.83	**28.57**	27.05	25.00	24.75	**21.30**	21.28	**30.19**
EBM												
BASE	18.56	17.33	**16.00**	11.20	14.95	**15.69**	18.56	17.33	**16.00**	11.20	14.95	**15.69**
CPT-only	**18.92**	**18.18**	15.19	11.77	14.64	14.04	19.13	**19.05**	15.87	11.77	14.29	13.95
MEDVOC	18.09	**18.18**	14.63	11.91	**15.31**	14.74	18.60	18.65	15.38	12.50	15.83	14.29
MEDVOC-LLM	18.77	17.20	14.55	12.50	14.45	15.27	**19.27**	17.54	14.71	**12.90**	**17.55**	14.81
ScafFix	18.67	16.67	15.39	**13.80**	15.27	15.63	18.65	17.40	14.71	11.77	15.47	14.71
BioASQ-M												
BASE	**28.50**	28.00	27.27	**21.53**	**26.71**	27.27	**28.50**	28.00	27.27	**21.53**	**26.71**	27.27
CPT-only	27.22	28.57	27.91	20.31	24.44	27.91	26.73	**29.41**	27.91	19.80	25.20	**28.57**
MEDVOC	24.19	26.32	23.26	16.53	23.08	26.09	24.47	28.07	21.74	16.06	24.19	24.57
MEDVOC-LLM	24.50	28.00	25.00	18.90	23.76	26.38	25.15	27.59	25.00	16.85	23.17	24.78
ScafFix	26.16	**30.00**	**29.79**	21.17	22.84	**28.07**	26.00	28.57	**28.57**	19.53	24.10	26.49
BioASQ-S												
BASE	32.12	35.72	26.32	**20.59**	**33.33**	**27.92**	32.12	35.72	26.12	20.53	**33.33**	**27.92**
CPT-only	33.09	29.35	25.53	19.64	23.65	24.12	**33.30**	30.33	29.22	20.00	30.30	21.92
MEDVOC	32.34	30.26	29.41	20.29	29.41	22.47	32.26	27.40	28.17	18.18	26.09	23.53
MEDVOC-LLM	32.75	34.52	30.30	29.29	33.33	24.47	32.40	30.33	**29.41**	20.29	27.03	21.53
ScafFix	**33.34**	**36.44**	**31.25**	19.05	27.03	24.00	32.88	**36.44**	29.29	**22.22**	**33.33**	24.00

Table 11: Performance comparison in terms of Rouge-L (R-L) between the two continual pretraining strategies of 'End-to-End' and 'Two-Stage' on Llama-2 7B model.

erence Summary (i.e., Difficult-RS Concentration), fraction of novel unigrams in reference summary (i.e., Novel-RS Concentration) and content overlap (measured by the standard metric Rouge-LCS) between source document and reference summary (i.e., Rouge-LCS overlap between source and reference). The differences between the two subsets in terms of these characteristics are compared in Table 10.

In terms of understanding the error, we made a key observation. The instances where BASE has better Rouge-LCS than ScafFix (Subset-1) have lesser values of Difficult-RS Concentration as well as Novel-RS Concentration, but higher values of Rouge-LCS overlap compared to Subset-2. Thus, for the instances that are less novel, have less OOV concentration, and are easier to infer, ScafFix is less helpful. This result from error analysis is also in line with the comparison of fine-grained settings (Table 6) reported in the main text, where we see general improvements for both Llama-2 and Llama-3.1 models for instances with higher OOV and higher novelty settings.

C Human Evaluation

We conducted our survey on Prolific platform[4] where we hired 9 medical experts from the platform across globe. All the annotators were shown 10 random samples from a pool of 30 summaries where the order of summaries was randomized and blinded (annotators have no idea which summary came from which model). The median time to complete was set at 30 mins and the annotators were paid at the rate of 8 UK pounds per hour based on the amount of time they took. We did not collect any PII from the participants explicitly other than what was provided by the platform. The task was conducted using Google Forms, with participants being shown a consent notice beforehand. The results are shown in Figure 3.

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

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

Annotation Guidelines. The annotations were assessed across three key dimensions as outlined by Fabbri et al. (2021): **Coherence**, **Relevance**, and **Factual Consistency**.

Coherence evaluates the structural integrity of the summaries, focusing on whether the sentences are logically connected and contextually aligned. **Relevance** measures the informativeness of the

[4]https://www.prolific.com/

	End-To-End						Two-Stage					
	$\text{Test}_{\text{Full}}$	$\text{Difficult}_{\text{SD}}$	$\text{Difficult}_{\text{RS}}$	Novel_{RS}	All_{SD}	All_{RS}	$\text{Test}_{\text{Full}}$	$\text{Difficult}_{\text{SD}}$	$\text{Difficult}_{\text{RS}}$	Novel_{RS}	All_{SD}	All_{RS}
PubMedQA												
BASE	**28.10**	24.46	26.87	18.87	**28.89**	29.21	**28.10**	24.46	26.87	18.87	29.23	32.10
CPT-only	26.62	24.52	**27.08**	21.07	25.93	26.74	27.05	**27.71**	26.97	19.64	**26.71**	28.65
MEDVOC	27.86	24.68	26.23	21.87	25.97	**31.12**	27.17	25.76	**27.12**	19.25	28.06	29.62
MEDVOC-LLM	27.69	**26.51**	26.67	21.74	25.53	30.43	27.15	25.19	26.51	19.86	28.45	28.51
ScafFix	27.67	23.85	25.00	**22.22**	25.00	29.16	27.25	26.37	26.42	**21.16**	28.16	**28.41**
EBM												
BASE	20.04	15.79	14.54	11.37	15.62	15.18	20.04	15.79	14.54	11.37	15.62	15.18
CPT-only	20.13	16.98	17.15	13.33	13.79	16.67	20.20	17.24	14.94	11.90	14.81	13.82
MEDVOC	20.31	17.93	16.03	12.17	13.56	16.87	20.45	**17.65**	15.56	13.34	**16.36**	**15.78**
MEDVOC-LLM	20.75	**18.39**	16.40	13.04	**16.13**	**17.05**	20.42	**17.65**	15.63	**14.82**	15.79	15.38
ScafFix	**20.79**	16.18	**17.17**	**13.80**	15.39	14.29	**20.50**	17.07	**16.85**	13.04	16.13	14.76
BioASQ-M												
BASE	**29.28**	28.57	26.67	**21.51**	25.54	23.77	**29.28**	28.57	26.67	**21.51**	**25.53**	23.77
CPT-only	27.56	28.57	26.09	19.84	25.54	23.53	27.25	28.57	**29.79**	19.14	24.56	**27.73**
MEDVOC	27.71	28.57	27.27	20.46	24.39	25.86	27.18	30.00	26.93	19.33	24.56	26.84
MEDVOC-LLM	27.45	29.17	27.27	19.05	25.00	26.09	27.92	**30.43**	29.03	20.97	25.00	27.36
ScafFix	28.91	**30.00**	**27.59**	21.23	**25.89**	**26.45**	27.86	28.13	28.57	20.94	24.10	27.38
BioASQ-S												
BASE	35.25	27.03	29.60	18.18	30.15	**27.92**	35.25	27.03	29.60	18.18	30.15	27.92
CPT-only	37.60	**42.42**	30.04	18.67	**42.02**	31.58	36.01	29.41	24.70	17.14	31.79	28.57
MEDVOC	37.00	31.25	27.03	20.29	30.38	28.92	37.01	32.43	29.79	18.18	32.43	31.11
MEDVOC-LLM	27.43	35.09	29.34	20.29	33.00	30.37	**37.15**	**37.84**	**32.89**	20.29	31.01	35.04
ScafFix	**37.22**	37.04	**33.34**	**20.34**	35.91	**35.02**	36.70	34.15	32.26	**20.90**	**33.74**	**33.35**

Table 12: Performance comparison in terms of Rouge-L (R-L) between the two continual pretraining strategies of 'End-to-End' and 'Two-Stage' on Llama-3.1 8B model.

summaries, considering the provided query and the context source document to judge the relevance. **Faithfulness** examines the accuracy of the stated facts, figures, and numerical data within the summaries, ensuring they can be directly verified against the input source. Notably, even if a summary presents accurate information, it is considered factually inconsistent if the claims cannot be substantiated solely by the given input. Relevance and Coherence are judged on a scale of 1-5; whereas factual consistency is binary (Yes/No).

To make the guidelines clear, we provide few examples of positive and negative examples to make people aware of what is a relevant, coherent, and factually consistent document vs what is not. The pdf is present in the github codebase.

D Representative Examples

We present examples from PubMedQA and EBM datasets where vocabulary adaptation strategy ScafFix helped improve the performance over BASE in higher OOV concentration settings from Llama-2 models.

	End-To-End						Two-Stage					
	$\text{Test}_{\text{Full}}$	**$\text{Difficult}_{\text{SD}}$**	**$\text{Difficult}_{\text{RS}}$**	**Novel_{RS}**	**All_{SD}**	**All_{RS}**	**$\text{Test}_{\text{Full}}$**	**$\text{Difficult}_{\text{SD}}$**	**$\text{Difficult}_{\text{RS}}$**	**Novel_{RS}**	**All_{SD}**	**All_{RS}**
PubMedQA												
BASE	25.40	**27.04**	**28.32**	**19.48**	22.97	29.29	25.40	**27.04**	**28.32**	19.48	22.97	**29.29**
CPT-Only	25.00	25.00	25.23	19.14	24.64	27.80	**25.51**	25.28	26.67	**19.84**	**24.81**	28.37
MEDVOC	24.49	24.09	26.67	17.56	23.88	28.27	24.64	24.99	25.01	18.85	22.79	26.49
MEDVOC-LLM	**25.45**	25.05	24.58	19.00	24.19	**29.62**	24.47	24.06	24.25	19.20	23.49	28.10
ScafFix	25.40	24.04	25.27	18.67	**24.62**	27.92	24.56	24.78	24.85	17.79	23.14	28.29
EBM												
BASE	**17.43**	**17.86**	**17.06**	11.49	**16.87**	12.33	**17.43**	**17.86**	**17.06**	11.49	**16.87**	13.33
CPT-Only	17.15	17.70	16.49	**12.12**	15.85	**13.79**	17.39	17.82	15.59	11.32	14.29	12.77
MEDVOC	17.63	16.09	14.04	11.24	15.46	12.99	17.10	15.83	14.61	**13.01**	15.38	**13.70**
MEDVOC-LLM	17.19	17.40	14.77	11.11	15.19	13.33	16.85	16.22	15.48	12.66	15.19	12.50
ScafFix	16.80	17.19	14.98	11.19	15.22	13.11	17.21	16.87	13.30	11.99	13.33	12.90
BioASQ-M												
BASE	**25.97**	**28.21**	**25.64**	**21.88**	**26.23**	**26.67**	**25.97**	**28.21**	**25.64**	**21.88**	**26.23**	**26.67**
CPT-Only	22.73	25.64	25.00	21.05	23.38	23.53	23.53	27.03	24.56	21.54	23.26	24.56
MEDVOC	21.54	23.68	17.02	16.67	22.20	21.43	21.95	25.30	15.38	17.02	22.95	22.22
MEDVOC-LLM	22.73	24.69	20.00	16.18	23.08	22.73	22.64	26.67	19.23	17.91	24.00	22.86
ScafFix	22.73	24.39	19.35	18.52	22.58	23.81	20.69	23.53	17.02	16.33	22.22	21.28
BioASQ-S												
BASE	**35.44**	**53.59**	**40.00**	**28.57**	**40.00**	**40.00**	**35.44**	**53.59**	**40.00**	**28.57**	**40.00**	**40.00**
CPT-Only	28.57	31.01	23.58	22.22	28.57	25.81	29.63	34.14	28.57	21.15	32.26	28.57
MEDVOC	28.57	31.16	26.32	19.15	27.03	27.78	27.78	29.61	25.93	19.61	28.00	23.81
MEDVOC-LLM	28.17	35.19	27.91	20.22	35.90	29.63	27.27	34.22	26.09	20.51	30.30	26.67
ScafFix	30.00	33.52	29.63	25.00	32.00	33.33	24.69	31.95	24.24	20.00	22.22	25.00

Table 13: Performance comparison in terms of Rouge-L (R-L) between the two continual pretraining strategies of 'End-to-End' and 'Two-Stage' on Mistral 7B model.

	End-To-End						Two-Stage					
	$\text{Test}_{\text{Full}}$	**$\text{Difficult}_{\text{SD}}$**	**$\text{Difficult}_{\text{RS}}$**	**Novel_{RS}**	**All_{SD}**	**All_{RS}**	**$\text{Test}_{\text{Full}}$**	**$\text{Difficult}_{\text{SD}}$**	**$\text{Difficult}_{\text{RS}}$**	**Novel_{RS}**	**All_{SD}**	**All_{RS}**
PubMedQA												
BASE	24.00	22.22	24.14	18.82	23.74	**30.00**	24.00	22.22	24.14	18.82	23.74	**30.00**
CPT-only	24.65	**27.85**	25.39	17.50	**24.85**	25.81	24.73	23.88	24.78	17.65	**25.41**	25.64
MEDVOC	24.59	23.08	**25.93**	18.37	23.21	26.32	24.89	25.45	26.20	18.18	23.57	25.00
MEDVOC-LLM	**25.00**	25.00	25.90	**19.50**	24.83	26.67	**25.00**	**25.00**	**27.06**	16.87	23.47	27.60
ScafFix	24.66	25.58	25.65	18.60	23.81	25.54	**25.00**	**25.00**	26.32	**18.87**	25.00	26.32
EBM												
BASE	17.52	16.49	14.04	11.43	16.07	14.46	17.52	16.49	14.04	11.43	16.07	14.46
CPT-only	17.60	15.38	15.17	11.86	16.42	13.91	**17.80**	14.93	14.89	11.51	14.77	13.95
MEDVOC	17.54	16.13	16.13	12.31	16.67	15.62	17.43	15.87	15.47	**12.50**	**16.75**	15.15
MEDVOC-LLM	17.34	16.13	14.70	12.60	16.16	15.15	17.65	**17.93**	15.39	11.76	16.39	15.15
ScafFix	**17.78**	**17.24**	**17.64**	**13.33**	**16.98**	**17.58**	17.67	17.24	**16.67**	12.04	16.07	**17.14**
BioASQ-$\mathcal{M}$												
BASE	**25.45**	30.00	**27.27**	19.67	26.09	**27.73**	**25.45**	**30.00**	**27.27**	19.67	**26.09**	**27.73**
CPT-only	23.88	28.57	26.10	18.52	25.25	26.90	24.56	28.57	25.00	17.86	25.35	26.90
MEDVOC	25.00	**30.20**	26.67	**20.00**	**26.67**	26.85	24.49	28.57	25.00	19.05	24.70	26.32
MEDVOC-LLM	24.30	28.57	25.71	19.23	25.35	27.27	24.34	28.57	26.67	**20.00**	25.00	26.80
ScafFix	24.49	28.24	26.23	19.67	25.00	27.20	24.69	28.57	26.67	**20.00**	25.00	27.27
BioASQ-$\mathcal{S}$												
BASE	**36.19**	**45.31**	**45.24**	**28.57**	**41.89**	**43.17**	**36.19**	**45.31**	**45.24**	**28.57**	**41.89**	**43.17**
CPT-only	32.43	37.89	28.53	24.49	35.59	29.18	31.59	31.50	32.57	25.00	34.46	29.90
MEDVOC	31.91	35.56	32.46	24.62	32.95	31.29	31.59	33.69	31.86	25.00	33.81	30.25
MEDVOC-LLM	32.14	30.49	30.94	26.32	33.81	30.77	32.14	37.97	31.84	24.56	36.18	34.29
ScafFix	32.00	39.36	36.94	26.67	37.23	31.01	31.58	38.28	32.13	26.00	36.04	29.10

Table 14: Performance comparison in terms of Rouge-L (R-L) between the two continual pretraining strategies of 'End-to-End' and 'Two-Stage' on Qwen2 7B model.

Input
Query: What medication best prevents migraine in children? **PubMed Abstract**: In a prospective study we compared propranolol, placebo, and self-hypnosis in the treatment of juvenile classic migraine. Children aged 6 to 12 years with classic migraine who had no previous specific treatment were randomized into propranolol (at 3 mg/kg/d) or placebo groups for a 3-month period and then crossed over for 3 months. After this 6-month period, each child was taught self-hypnosis and used it for 3 months. Twenty-eight patients completed the entire study. The mean number of headaches per child for 3 months during the placebo period was 13.3 compared with 14.9 during the propranolol period and 5.8 during the self-hypnosis period. Statistical analysis showed a significant association between decrease in headache frequency and self-hypnosis training (P = .045). There was no significant change in subjective or objective measures of headache severity with either therapy.
Reference Summary (OOV Concentration: 17.72%)
A comparative randomized controlled trial with multiple crossovers involving 33 children found that a self-hypnosis placebo decreased mean headache frequency from 13.3 per 3-month interval to 5.8 (P=.045), but found propranolol no different than placebo. Propranolol was also studied in a 3-armed randomized controlled trial in comparison with flunarizine-a drug likely to be efficacious and placebo.
BASE Summary (Rouge-L: 24.10)
Self-hypnosis was associated with a significant decrease in headache frequency in a single randomized controlled trial of 28 children.
ScafFix Summary (Rouge-L: 40.00)
In a single randomized controlled trial with crossover design in 28 children, self-hypnosis reduced headache frequency from a mean of 13.3 attacks/month to 5.8 attacks/month vs. no change for placebo (P =.045).

Table 15: Example from EBM dataset comparing summary generated using vocabulary adaptation method ScafFix with BASE.

Input
Query: Should serum pancreatic lipase replace serum amylase as a biomarker of acute pancreatitis? **PubMed Abstract**: Serum pancreatic lipase may improve the diagnosis of pancreatitis compared to serum amylase. Both enzymes have been measured simultaneously at our hospital allowing for a comparison of their diagnostic accuracy. Seventeen thousand five hundred and thirty-one measurements of either serum amylase and or serum pancreatic lipase were made ... One thousand eight hundred and twenty-five patients had either elevated serum amylase and or serum pancreatic lipase. The medical records coded for pancreatitis in a further 55 whose enzymes were not elevated. Three hundred and twenty of these had radiological evidence of acute pancreatitis. Receiver operator characteristic analysis of the initial sample from patients received in the Emergency Department showed improved diagnostic accuracy for serum pancreatic lipase (area under the curve (AUC) 0.948) compared with serum amylase (AUC, 0.906, P<0.05) ...
Reference Summary (OOV Concentration: 47.09%)
It is concluded that serum pancreatic lipase is a more accurate biomarker of acute pancreatitis than serum amylase.
BASE Summary (Rouge-L: 29.17)
Serum pancreatic lipase may improve the diagnosis of pancreatitis compared to serum amylase. Both enzymes have been measured simultaneously at our hospital allowing for a comparison of their diagnostic accuracy.
ScafFix Summary (Rouge-L: 76.47)
Serum pancreatic lipase is a more sensitive and specific biomarker of acute pancreatitis than serum amylase.

Table 16: Example from PubMedQA dataset comparing summary generated using vocabulary adaptation method ScafFix with BASE. Interestingly, BASE model suffered from LEAD-bias (See et al., 2017) (picking sentences from the front of the document as the summary) and missed the biomarker word completely.