In-Domain Pre-Training Improves Clinical Note Generation from Doctor-Patient Conversations

Colin A. Grambow **Longxiang Zhang** **Thomas Schaaf**
3M Health Information Systems
cgrambow,lzhang28,tschaaf@mmm.com

Abstract

Summarization of doctor-patient conversations into clinical notes by medical scribes is an essential process for effective clinical care. Pre-trained transformer models have shown a great amount of success in this area, but the domain shift from standard NLP tasks to the medical domain continues to present challenges. We build upon several recent works to show that additional pre-training with in-domain medical conversations leads to performance gains for clinical summarization. In addition to conventional evaluation metrics, we also explore a clinical named entity recognition model for concept-based evaluation. Finally, we contrast long-sequence transformers with a common transformer model, BART. Overall, our findings corroborate research in non-medical domains and suggest that in-domain pre-training combined with transformers for long sequences are effective strategies for summarizing clinical encounters.

1 Introduction

Necessitated by electronic health records (EHR), physicians spend a large amount of time on documentation work (Sinsky et al., 2016), which contributes significantly to burnout (Wright and Katz, 2018; Kumar and Mezoff, 2020), may result in lower job satisfaction (Shanafelt et al., 2016), and can even increase the likelihood of errors and reduce the quality of patient care (Panagioti et al., 2017). To alleviate some of the burden on physicians, medical scribes are often used to summarize recordings or transcriptions of doctor-patient conversations into clinical notes. While this essential, yet tedious process may enable more effective clinical care, it shifts the burden onto medical scribes. Furthermore, the continued reliance on human experts is expensive and only scalable to a limited degree.

Natural language generation models, such as the ones developed in this paper, have the potential to significantly reduce the documentation burden by providing suggested clinical notes to physicians or scribes nearly instantaneously. While still somewhat error-prone and not yet fully automated, these models are able to focus on much of the relevant information in doctor-patient conversations and distill it into a human-readable format for further review by trained medical professionals.

Pre-trained transformer models have revolutionized the field of natural language processing (Radford et al., 2019; Devlin et al., 2019; Lewis et al., 2020) and have already been applied to various medical tasks (Lee et al., 2019; Li et al., 2020; Zhang et al., 2021; Yalunin et al., 2022). Nonetheless, medical conversation summarization continues to present challenges due to its idiosyncrasies, foremost of which is the requirement to contain all relevant medical information rather than summarizing every part of a conversation. Additionally, specialized medical vocabulary renders the use of conventional pre-trained models difficult.

Additional phases of in-domain pre-training have shown to be useful across a wide variety of domains and tasks (Gururangan et al., 2020), but limited work has been done on in-domain pre-training using unlabeled doctor-patient conversations. To address this, we leverage a doctor-patient conversation dataset described in Section 3 to investigate two different pre-training methodologies using BART, LED, and DialogLED transformer models (Section 4). We fine-tune all models on a subset of medical conversations with human-written summaries (Section 4.2) and contrast them with a baseline of models that are not pre-trained in the medical domain using several different evaluation methods, including a transformer-based model for clinical concept extraction (Sections 4.3 and 5). We show that our methods improve the performance on the medical summarization task and also evaluate the additional benefit of using models designed to work with long sequences (Section 6).

Proceedings of the First Workshop on Natural Language Generation in Healthcare (NLG4Health), pages 9 - 22

2 Related Work

Medical summarization Recent research has devoted significant attention to the problem of summarizing medical encounters and documents in an automated fashion. Finley et al. (2018) describe a fully automated medical scribe using a combination of RNN and rule-based approaches to automatically recognize speech, convert it into a transcript, extract the relevant information, and convert it to a report. However, they omit any examples and results and mention that the scribe is still limited in its utility.

Since then, several deep learning approaches have been developed to summarize doctor-patient conversations. Joshi et al. (2020) develop a modified pointer-generator (PG) network to summarize local snippets. Furthermore, they explicitly model negation, which can cause difficulties for automatic approaches. Interestingly, they report that transformer models did not work well, which is contrary to the findings in a lot of subsequent research. Yim and Yetisgen (2021) also use a PG model to perform the similar task of sentence alignment and snippet summarization. Notably, they achieve good results using only a very small dataset. Krishna et al. (2021) take on the challenging task of generating complete clinical summaries (SOAP notes) using various LSTM, PG, and transformer models. They extract important utterances, cluster them, and then generate single-sentence summaries of each cluster. Enarvi et al. (2020) use a large dataset of doctor-patient conversations generated using automatic speech recognition to train a combined transformer-PG model from scratch. They are able to handle somewhat longer input because they do not rely on pre-trained transformer models. As an alternative approach to handle long conversations, Zhang et al. (2021) use a pre-trained BART model with a two-stage chunking approach to generate summaries for a section of the clinical notes.

Related to the summarization of doctor-patient conversations, other research has explored the summarization of clinical notes and clinical history. Zhang et al. (2018) use a PG network to summarize radiology findings and found that incorporating additional information in the form of background information about the patient improves the results. Yalunin et al. (2022) construct a model using a Longformer encoder with a BERT decoder to generate parts of discharge notes from the patient history. They pre-train BERT and Longformer on domain-specific data and create a custom tokenizer, which yields strong results.

Domain shift An intrinsic problem with using pre-trained models is that the domain of the pre-training data is often significantly different from that of the target medical domain. PG networks during fine-tuning can be helpful because they are able to copy words from the new vocabulary, but starting from a model in a domain that is closely related to that of the fine-tuning task would provide additional benefit. Gururangan et al. (2020) show that a second round of pre-training in a domain related to the fine-tuning task can provide significant benefit even if the continued pre-training only uses the unlabeled training set for a given task. They investigate this across a broad range of domains and classification tasks. Similarly, Hsu et al. (2021) find that in-domain pre-training improves learning speech representations. Zhong et al. (2021) show that improved summarization results are possible by continuing pre-training in the (non-medical) conversation domain. As already mentioned previously, Yalunin et al. (2022) use in-domain pre-training very successfully for generating discharge notes from patient histories.

Instead of pre-training all model parameters in the new domain, there has been some investigation into learning small extension modules instead, which can be helpful if there are limited pre-training data or if complete model training is too costly. Tai et al. (2020) adapt BERT to the medical domain by creating an additional vocabulary and adding a corresponding embedding layer. They compose their extension module as a weighted summation of the embedding vectors from the original and the extension layers and demonstrate that this method is very effective at adapting to the new domain.

3 Dataset

The dataset we use has already been described by Zhang et al. (2021) and is composed of 83 605 clinical encounters involving doctors from many different specialties, patients, and potentially other speakers, *e.g.*, nurses and caregivers. For each encounter, we use the de-identified doctor-patient conversation transcribed by a human. The median number of tokens in a conversation is 2040 (using the BART byte-pair encoding from Lewis et al., 2020), and there are a total of 203M tokens in the

entire dataset.

Annotations are available for a subset of 1342 conversations in the form of medical summaries across internal medicine and primary care specialties. Each conversation was summarized by multiple professional medical scribes into several sections, of which we only use the History of Present Illness (HPI) section for this paper due to its complexity and because it is usually written in complete sentences. There are an average of 17 reference summaries per doctor-patient conversation. The median conversation length in the summarization subset is 1334 tokens for a total of 2.5M tokens. The 95$^{\text{th}}$ percentile corresponds to approximately 5120 tokens, which is the length limit we use for our long-sequence models.

For pre-training, we exclude the entire subset of data that we have summaries for in order to avoid any data leakage and potentially biased results. In addition, we split off a random 5% of the remaining conversations as the validation dataset to monitor during pre-training.

For fine-tuning, we attempt to remove poor summaries for a given conversation using an in-house rule-based system to extract medical concepts from the training summaries and only keeping the summary with the most concepts. Even though this results in fewer labeled data, we have not observed a significant drop in performance. Nonetheless, we keep all reference summaries for the test data. After splitting and removing extraneous summaries from training and validation data, we end up with 939, 201, and 202 conversations; and 939, 201, and 3450 summaries in the training, validation, and test sets, respectively.

4 Methods

All methods are based on pre-training and/or fine-tuning of BART (Lewis et al., 2020), Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020), and DialogLED (Zhong et al., 2021). BART is a pre-trained encoder-decoder transformer model designed for fine-tuning on text generation tasks, such as summarization. However, it can only encode up to 1024 tokens in both its encoder and decoder, which is less than the median sequence length in our fine-tuning dataset. LED and DialogLED can handle significantly longer input sequences (we use 5120 and 1024 tokens for their encoders and decoders, respectively) by employing a combined global and local attention mechanism which scales linearly with sequence length. The LED architecture is almost identical to that of BART except that the position embeddings of BART are copied 16 times to enable longer input. The parameters of LED are initialized from BART and no additional pre-training was done. DialogLED is initialized from LED and further pre-trained on long dialog data using a window-based denoising task specifically designed for conversations, which results in significant improvement for long-dialog summarization.

We initialize and train all of our models using the pre-trained BART, LED, and DialogLED models available in the Hugging Face Transformers library (Wolf et al., 2020). We use the corresponding tokenizers (all of which use the BART/GPT-2 byte-pair encoding with a vocabulary size of 50 265), but we add additional speaker tokens, *e.g.*, `[DR]:`, `[PT]:`, etc. We investigate both the base and large models (140M vs. 400M parameters, respectively). Except for the additional position embeddings, all base models and all large models have the same number of parameters.

4.1 Pre-training

We investigate two types of pre-training with doctor-patient conversations: BART-style denoising using the entire input as described by Lewis et al. (2020) and DialogLED-style window-based denoising as described by Zhong et al. (2021). We found that sentence and turn permutation are always detrimental to the downstream summarization of our doctor-patient conversations as measured by a decrease in ROUGE scores, so we only perform text infilling for BART-style pre-training and we only use speaker masking, turn splitting, turn merging, and text infilling for window-based pre-training. The other denoising hyperparameters are identical to those used in the original papers. For BART-style denoising, we discovered that it is beneficial to allow the attention mechanism to attend to the additional padding tokens that are added as a result of the text infilling. We hypothesize that this could imply that adding noise to the entire input is too "difficult" of a pre-training task so that some additional information is necessary in the form of the padding tokens, but we leave the further investigation of this observation to a future study.

For all models, we split the conversations into chunks of 1024 tokens for BART-style pre-training,

and we simply truncate long conversations at 5120 tokens for window-based denoising with LED and DialogLED. The number of epochs that each model is pre-trained for is chosen to achieve optimal performance on the downstream summarization task. For the large models, this results in less than one full pass across the pre-training dataset being required. All of our pre-training hyperparameters are shown in the Appendix in Table A.1.

4.2 Fine-tuning

Our fine-tuning task is training a text generation model to summarize doctor-patient conversations into coherent HPI summaries containing all relevant medical information. As with pre-training, we use a decoder sequence length of 1024 tokens and encoder sequence lengths of 1024 tokens for BART and 5120 tokens for LED and DialogLED. 5120 tokens corresponds to the 95$^{\text{th}}$ percentile of conversations in the summarization dataset, which allows us to encode the full length of the majority of the conversations when using LED and DialogLED. Other than that, we maintain consistency across all other fine-tuning hyperparameters (see Table A.2) for all of our models. We train for a maximum of 30 epochs with a batch size of 8 and evaluate every 50 steps. We perform evaluation by using the validation data input to generate text using beam search and monitor the geometric mean of ROUGE-1 F1, ROUGE-2 F1, and ROUGE-L F1 scores on the validation data. We stop training if the validation score has not improved over the last five evaluation calls and save the best model checkpoint.

4.3 Evaluation

In order to rapidly estimate performance across all reference and generated summaries, we employ several automatic evaluation methods. In addition to ROUGE and UMLS concept-based evaluation, which have been used previously in the literature, we also suggest a named entity recognition model as a second form of concept-based evaluation due to the ease of fine-tuning such a model on publicly available data.

4.3.1 ROUGE

We use the `rouge-score` package[1] to compute ROUGE scores, which aims to replicate results from Lin (2004). While there are some issues with using ROUGE for abstractive summarization (Kryscinski et al., 2020), especially with regard to hallucination (Maynez et al., 2020), it is a useful metric to assess the degree of overlap between reference and generated summary. As there are multiple reference summaries per conversation in the fine-tuning test set, we first compute the ROUGE scores of a generated summary with all of its corresponding reference summaries for a single doctor-patient conversation and then average each score. To obtain an aggregate ROUGE score, we can then average the scores across all conversations.

4.3.2 Clinical concepts

As ROUGE measures word overlap indiscriminately, it takes into account unimportant words and is not as suitable for measuring semantic overlap. Therefore, it is beneficial to quantify additional metrics that are not as prone to these issues and focus more on the relevant medical content of a summary.

UMLS concept extraction The methodology described in this paragraph is largely identical to the evaluation described by Zhang et al. (2021). The Unified Medical Language System (UMLS) (Bodenreider, 2004) is a large database of medical concepts and relations between them. We use the approximate string matching algorithm implemented in QuickUMLS (Soldaini and Goharian, 2016) to extract strings from our summaries and match them to concepts in the UMLS database. However, this approach sometimes mislabels irrelevant strings as medical concepts. To mitigate this somewhat, we first aggregate and filter concepts from all reference summaries for a given conversation by only keeping a concept if it occurs in at least three reference summaries or if it occurs in all reference summaries if there are fewer than three. We then extract the UMLS concepts for the generated summary and compute precision, recall, and F1-score. Aggregate scores are averaged across all conversations.

Transformer-based clinical concept extraction (NER) To further deal with the limitations of QuickUMLS, such as the extraction of irrelevant strings from a summary, we train a deep learning model to extract clinical concepts instead. For this, we follow the clinical concept extraction approach by Yang et al. (2020). We use their RoBERTa (Liu et al., 2019) model pre-trained on MIMIC-III clinical notes (Johnson et al., 2016) to fine-tune a named entity recognition (NER) model on the i2b2 2010 dataset (Uzuner et al., 2011), which is a large col-

[1] https://github.com/google-research/google-research/tree/master/rouge

lection of clinical notes annotated with three types of medical concepts.

First, we mostly reproduce the strong classification performance reported by Yang et al. (2020) using the conventional i2b2 2010 train-test split and a conditional random field layer on top of the transformer model (see Table B.1 in the Appendix). After verifying that this approach is successful, we train our final clinical concept extraction model on all i2b2 2010 data for use on our summaries.

To automate the NER-based concept evaluation, we map the extracted entities to UMLS concept unique identifiers (CUIs) using QuickUMLS (there are frequently multiple CUIs per entity) and drop any entities that cannot be mapped. We combine entities that are of the same type (as predicted by the NER model) and have overlapping sets of UMLS CUIs. Similar to the QuickUMLS-only approach, we only keep reference summary entities if they occur in at least three reference summaries for a given conversation. Finally, we compute precision, recall, and F1-score. For this, we define a true positive as a concept extracted from the generated summary where its predicted type matches that of a concept extracted from the reference summaries and there exists an intersection between the sets of UMLS CUIs corresponding to the concepts. False positives and false negatives are defined accordingly.

5 Experiments

We establish baselines by fine-tuning base and large versions of vanilla BART, LED, and DialogLED models on the doctor-patient conversation summarization dataset as described in Section 4.2, *i.e.*, using the versions of those models that are pre-trained as described in their original papers. To assess whether a second round of pre-training on in-domain data is beneficial, we continue pre-training the models on our doctor-patient conversation dataset as described in Section 4.1 followed by fine-tuning on the summarization dataset.

For BART, window-based denoising results in a negative impact on ROUGE scores, so we only investigate normal text infilling denoising, whereas for LED and DialogLED, we consider both BART-style text infilling and window-based denoising. The results of performing all types of evaluation described in Section 4.3 on the summarization test set are shown in Table 1. Furthermore, we report the median length of generated summaries in Table 2.

6 Qualitative Analysis

In-domain pre-training Across all models and pre-training objectives, ROUGE *F1* scores always improve with additional in-domain pre-training (Table 1), clearly indicating that pre-training leads to improved overlap between the generated and reference summaries. For the sake of completeness, it should be mentioned that we find that ROUGE *precision* generally decreases with increasing sequence length (Table 2) whereas ROUGE *recall* generally increases; however, we see no such correlation for ROUGE F1 so that we continue to use ROUGE F1 for the discussion here. The full evaluation results, including precision and recall can be found in Table C.1 in the Appendix. There exists some research into removing the length bias from ROUGE score calculations (*e.g.*, Sun et al., 2019), but this is out of scope for our current study.

Overall, we find that pre-training LED with the window-based denoising task leads to the strongest models in terms of ROUGE scores. For LED-large, in-domain pre-training improves the summarization performance of doctor-patient conversations by 1.59 points for ROUGE-1, 1.13 points for ROUGE-2, and 1.10 points for ROUGE-L relative to the vanilla LED-large baseline (Table 1).

Similarly, in-domain pre-training almost always improves both of our concept-based evaluation metrics with the only noticeable outlier being BART-large. We note that we observe a slightly stronger dependence of precision and recall on sequence length (Table 2 and Appendix Table C.2) than with ROUGE. Nonetheless, in-domain pre-training leads to the best-performing models as measured by concept-based F1 scores even if the pre-trained version does not generate longer sequences on average.

Overall, we find that pre-training DialogLED with the BART-style text infilling task leads to the strongest models in terms of concept-based scores, which is contrary to the performance of DialogLED when measured with ROUGE. This could imply that while DialogLED generates extraneous text (also shown by its long generation length in Table 2) which results in lower ROUGE scores, it is better at generating the relevant medical concepts, which might make it more useful for medical summarization.

Comparing across denoising tasks used for pre-training, there seems to be no significant difference in terms of ROUGE between BART-style

Model	Pre-train	ROUGE F1			UMLS F1	NER F1
		R-1	*R-2*	*R-L*		
BART-base	—	35.19	13.28	25.43	24.27	36.22
BART-base	BART	36.83	**14.13**	26.53	28.63	40.24
LED-base	—	36.01	13.49	25.99	26.40	38.59
LED-base	window	**36.96**	14.12	**26.72**	27.45	39.96
LED-base	BART	36.65	13.60	26.31	28.47	41.39
DialogLED-base	—	36.07	13.14	25.13	31.22	42.66
DialogLED-base	window	36.85	13.79	25.74	31.62	41.87
DialogLED-base	BART	36.79	13.59	25.88	**33.33**	**42.72**
BART-large	—	38.25	14.78	26.65	35.19	47.52
BART-large	BART	38.47	14.89	**27.37**	27.77	43.45
LED-large	—	37.29	13.83	26.09	30.45	43.97
LED-large	window	**38.88**	**14.96**	27.19	32.03	46.78
LED-large	BART	38.07	14.56	26.82	35.33	47.15
DialogLED-large	—	37.04	13.74	25.55	32.36	47.23
DialogLED-large	window	37.26	14.15	25.73	34.05	45.86
DialogLED-large	BART	37.73	14.56	25.69	**38.90**	**51.57**

Table 1: Evaluation results on the summarization test set. In the "Pre-train" column, "BART" refers to BART-style pre-training without sentence permutation (text infilling across the entire input) and "window" refers to window-based denoising (without turn permutation). The metrics from left to right are ROUGE-1 F1, ROUGE-2 F1, ROUGE-L F1, QuickUMLS concept-based F1, and NER concept-based F1.

Model	Pre-train	Median summary length
BART-base	—	53
BART-base	BART	62
LED-base	—	62
LED-base	window	65
LED-base	BART	65
DialogLED-base	—	71
DialogLED-base	window	73
DialogLED-base	BART	71
BART-large	—	89
BART-large	BART	70
LED-large	—	98
LED-large	window	81
LED-large	BART	88
DialogLED-large	—	108
DialogLED-large	window	110
DialogLED-large	BART	115
Training set reference summaries		114
Test set reference summaries		81

Table 2: Median sequence length in number of tokens of generated summaries and of summaries in the training (with validation) and test data.

text infilling and window-based denoising, whereas concept-based scores improve with BART-style pre-training compared to window-based denoising. Even though Zhong et al. (2021) designed the window-based denoising task for conversation data, it seems that it is not always beneficial to use over more conventional pre-training. The most important thing is simply the process of pre-training on medical conversations itself, regardless of pre-training objective used.

One benefit of using automatic metrics is that they may quantify smaller improvements which would not be as visible with small-scale human evaluation. In particular, Table 3 compares example output from vanilla LED-large and from LED-large pre-trained on doctor-patient conversations using window-based denoising. While there are some differences in the output, it is not immediately evident that the output from the pre-trained model is better. Both models produce fluent summaries and include all of the important concepts mentioned in the reference summary. In the conversation, the patient incorrectly refers to the mitral valve as "microvalve", and no explicit mention of the correct term occurs (the relevant conversation snippet is shown in Table 4). Naturally, a trained

Source	Summary
Reference	The patient is a female presenting today for routine follow-up. She states that she is doing well and continues to take Fosamax as prescribed. She continues to experience pain to her back. She is also requesting to have her heart checked as she was diagnosed with mitral valve prolapse several years ago.
BART-large (vanilla)	The patient is a female presenting to the clinic today for a follow up visit concerning her hypercholesterolemia. She reports that she has been on Fosamax for at least 4 years. She has been taking it weekly for the past 9 years. Her last bone density test was in June of last year. She is still taking Prilosec every day.
LED-large (vanilla)	The patient is a female presenting to the clinic today for a follow-up visit. She agreed to a virtual scribe. Back Pain - She has noticed that her back has been bothering her for the last month. It does not hurt to push, but it is bothersome. She takes Tylenol or Advil if it is really annoying. She has been taking Fosamax for the past 4 years. She would like to have a stress test of her arteries to see if they are strong. She is currently taking Prilosec every day. Heart Failure - Her last heart exam was 5 years ago. She had a microvalve prolapse at that time.
LED-large (window-based pre-trained)	The patient is a female presenting to the clinic today for a follow up visit. She has a history of hyperlipidemia and hypercholesterolemia. She states that she has been taking Fosamax for the past 4 years. She reports that her back has been bothering her for the last month. She denies any fractures or fractures in her bones. She is taking Prilosec every day. She would like to know how her arteries are doing and if she needs more vitamin D. She also wants to know if she has a microvalve prolapse.

Table 3: Comparison of reference summary and several generated summaries for a conversation with 2088 tokens from the test set.

medical scribe uses the correct term in the summary, whereas the LED models are not able to perform this line of complex reasoning without additional information, so they copy the term used by the patient. The vanilla LED model makes another error by stating that the patient takes Tylenol or Advil; however, the doctor is the one to suggest this in the conversation, the patient never made such a statement. A small error also occurs in the pre-trained LED model, which mentions that the patient is inquiring about vitamin D, but this is also something said by the doctor, not the patient.

Long conversations Vanilla BART-large is a strong baseline that cannot always be outperformed by the long-sequence models (Table 1). In fact, DialogLED is noticeably weak in terms of ROUGE which might imply that a non-trivial amount of information was lost during the first round of continued pre-training (on non-medical long-dialog data). Such a direct comparison between DialogLED and BART is possible because DialogLED is a further pre-trained version of LED, which is itself initialized from BART. However, as mentioned earlier, concept-based evaluation of DialogLED shows strong performance, indicating that ROUGE alone may not be sufficient for quantifying the utility of a model.

For a different reason than DialogLED, vanilla LED is also a weak baseline. We observed difficulty during fine-tuning of vanilla LED on our small dataset and hypothesize that this could be a result of non-ideal initialization of its copied position embeddings (Beltagy et al., 2020). As the position embeddings for positions greater than 1024 never underwent their own additional pre-training, their parameters are not necessarily optimal at the start of fine-tuning. However, in-domain pre-training results in a suitable initialization for the position embeddings prior to fine-tuning, which manifests itself in good performance compared to pre-trained BART after fine-tuning. Still, LED is not significantly better than BART after in-domain pre-training even though it can process much longer input. One possible reason is that most of the relevant information for the HPI section might be contained at the beginning of long conversations.

[PT]:	And, uh, I want to have my heart checked out because my heart, um, I think it was five years ago -
[DR]:	Um-hum.
[PT]:	We did it. I have a microvalve prolapse.
[DR]:	Um-hum.
[PT]:	But they said at that time it wasn't that bad -
[DR]:	Um-hum.
[PT]:	But, um, I feel like I need to check up on that again. And can they do the, uh, also can they do the arteries? Can they check your arteries?
[DR]:	They do that with the stress test. The stress test is a way of, um, the stress test has a way of looking at the arteries. You don't want to actually have the dye put in your arteries because that is dangerous.

Table 4: Snippet of the conversation corresponding to Table 3 revolving around the heart valve prolapse.

Another reason is that our median conversation length in the fine-tuning dataset (see Section 3) is not much longer than the maximum input size BART can process, so there may not be enough long conversations for the difference in models to make a large difference.

If we bin the conversations by their number of tokens and compare BART-large to LED-large, we observe less of a drop in ROUGE for longer conversations with LED-large than with BART-large (Figure 1), suggesting that LED does extract additional useful information from long inputs. The improved performance on long conversations with LED-large is even more evident when analyzing the concept-based metrics across different conversation lengths as shown in Figure 2. LED-large is very effective at extracting relevant concepts from long conversations.

The example in Table 3 corroborates this finding: The summary generated by BART-large fails to mention the back pain and heart valve prolapse, whereas LED-large correctly includes both of these concepts. Both concepts are only mentioned in the latter half of the conversation, which, with a length of 2088 tokens, is significantly longer than the maximum BART sequence length. Unrelated to conversation length, the BART-large model is seemingly confused by the duration for which the patient has been taking Fosamax. However, the BART output is actually more accurate than the LED output, which states a duration of four years. In the conversation, the doctor is briefly confused about the Fosamax duration and initially assumes "at least four years", but then corrects that estimate to "at least nine years" over the course of several subsequent sentences.

Generated summary lengths We can observe several trends in the lengths of generated summaries in Table 2. First, large models generate longer summaries than base models, and while good performance is possible using base models (Table 1), this might hint at an inadequate intrinsic capacity of small models to model complex abstractive summarization, suggesting that one would be better served by using the large models. Second, pre-trained base models generate longer summaries than their corresponding vanilla versions with the exception of DialogLED-base, which could be a result of it already having been pre-trained on long-dialog data. Interestingly, this effect seems to be reversed for the large models: pre-trained BART-large and LED-large generate shorter summaries than their vanilla versions while pre-trained DialogLED-large generates slightly longer text. Third, DialogLED always generates the longest summaries compared to BART and LED even if these have been pre-trained on in-domain data. Again, this could be due to the round of pre-training on (non-medical) long-dialog data that DialogLED underwent.

On average, the generated summaries are shorter than those in the fine-tuning training set, although they happen to correspond well in length to those in the test set. As described in Section 3, the training set summaries are longer on average because they only contain the references with the most concepts extracted using our in-house rule-based system. Overall, these results indicate that there might be a need to bias the models toward longer generation length. However, we do not add any sort of length penalty here because our goal was to compare what the models learn in an unbiased fashion.

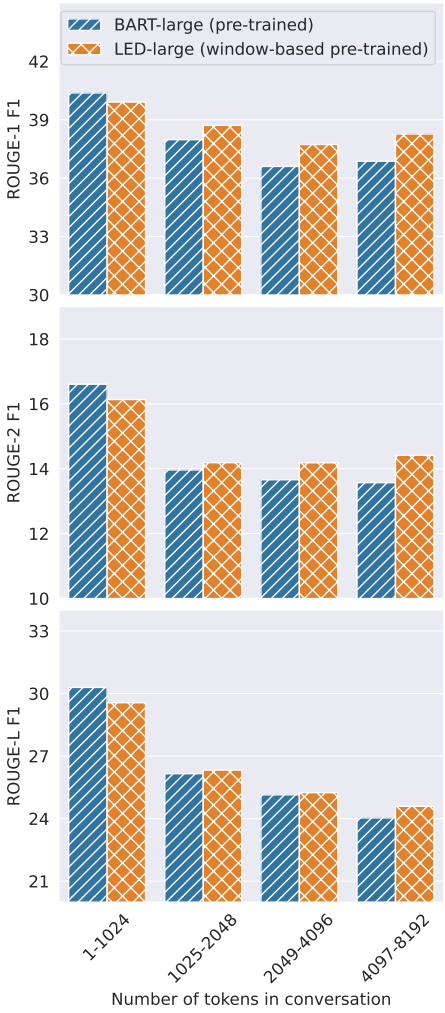

Figure 1: ROUGE score comparison binned by conversation length.

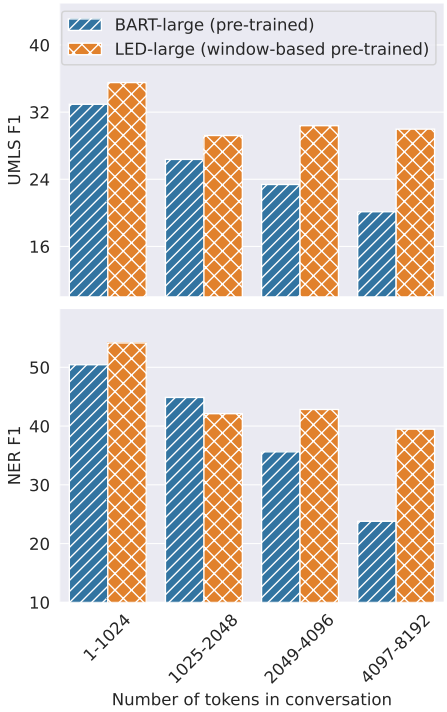

Figure 2: Concept-based score comparison binned by conversation length.

7 Conclusion

We showed that in-domain pre-training improves abstractive summarization of long doctor-patient conversations into HPI notes across several models based on the BART architecture and across two different pre-training objectives. To measure the improvement, we used conventional evaluation methods like ROUGE and UMLS concept-based evaluation and also trained a neural clinical concept extraction model to better extract relevant concepts. We also demonstrated the benefit of using models that can deal with long conversations intrinsically, especially for ensuring that relevant medical concepts are present.

While unlabeled doctor-patient conversations are a useful source of pre-training data, we hope to investigate additional types (*e.g.*, clinical notes) in the future. Similar research has already shown that other types of pre-training data can be very effective, *e.g.*, pre-training on patient histories (Yalunin et al., 2022) or pre-training on clinical notes for named entity recognition (Yang et al., 2020). Additionally, we can explore combining and contrasting our holistic pre-training approach with methods that only pre-train a small amount of additional parameters (Tai et al., 2020).

Lastly, given the varying lengths of generated summaries, we are considering methods to control generation length as another future research direction (Kikuchi et al., 2016).

Ethical Considerations

The models developed in this paper may omit important information or incorrectly include misleading details in the output they generate. Due to this, we stress the importance of not using the generated outputs unsupervised. In all cases, medical experts should review and edit the generated summaries. Nonetheless, we expect that our models can act as virtual assistants to alleviate some of the documentation burden.

The data used for pre-training and fine-tuning inherently contain sensitive medical information. To protect private health information, the data were manually de-identified by medical experts and no private information was used in the methods described in this paper.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *Computing Research Repository*, arXiv:2004.05150. Version 2.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–270.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 22–30, Online. Association for Computational Linguistics.

Gregory Finley, Erik Edwards, Amanda Robinson, Michael Brenndoerfer, Najmeh Sadoughi, James Fone, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. 2018. An automated medical scribe for documenting clinical encounters. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–15, New Orleans, Louisiana. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 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.

Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. 2021. Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training. *Computing Research Repository*, arXiv:2104.01027. Version 2.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. 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 4958–4972, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Gogi Kumar and Adam Mezoff. 2020. Physician Burnout at a Children's Hospital: Incidence, Interventions, and Impact. *Pediatric Quality & Safety*, 5(5):e345.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, page btz682.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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, Online. Association for Computational Linguistics.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. BEHRT: Transformer for Electronic Health Records. *Scientific Reports*, 10(1):7155.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Computing Research Repository*, arXiv:1907.11692.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Maria Panagioti, Efharis Panagopoulou, Peter Bower, George Lewith, Evangelos Kontopantelis, Carolyn Chew-Graham, Shoba Dawson, Harm van Marwijk, Keith Geraghty, and Aneez Esmail. 2017. Controlled Interventions to Reduce Burnout in Physicians: A Systematic Review and Meta-analysis. *JAMA Internal Medicine*, 177(2):195–205.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Tait D. Shanafelt, Lotte N. Dyrbye, Christine Sinsky, Omar Hasan, Daniel Satele, Jeff Sloan, and Colin P. West. 2016. Relationship Between Clerical Burden and Characteristics of the Electronic Environment With Physician Burnout and Professional Satisfaction. *Mayo Clinic Proceedings*, 91(7):836–848.

Christine Sinsky, Lacey Colligan, Ling Li, Mirela Prgomet, Sam Reynolds, Lindsey Goeders, Johanna Westbrook, Michael Tutty, and George Blike. 2016. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. *Annals of Internal Medicine*, 165(11):753–760.

Luca Soldaini and Nazli Goharian. 2016. QuickUMLS: a Fast, Unsupervised Approach for Medical Concept Extraction. In *Medical Information Retrieval (MedIR) Workshop, SIGIR 2016*.

Simeng Sun, Ori Shapira, Ido Dagan, and Ani Nenkova. 2019. How to compare summarizers without target length? pitfalls, solutions and re-examination of the neural summarization literature. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 21–29, Minneapolis, Minnesota. Association for Computational Linguistics.

Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending pretrained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online. Association for Computational Linguistics.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–556.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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. Association for Computational Linguistics.

Alexi A. Wright and Ingrid T. Katz. 2018. Beyond Burnout — Redesigning Care to Restore Meaning and Sanity for Physicians. *New England Journal of Medicine*, 378(4):309–311.

Alexander Yalunin, Dmitriy Umerenkov, and Vladimir Kokh. 2022. Abstractive summarization of hospitalisation histories with transformer networks. *Computing Research Repository*, arXiv:2204.02208.

Xi Yang, Jiang Bian, William R Hogan, and Yonghui Wu. 2020. Clinical concept extraction using transformers. *Journal of the American Medical Informatics Association*, 27(12):1935–1942.

Wen-wai Yim and Meliha Yetisgen. 2021. Towards automating medical scribing : Clinic visit Dialogue2Note sentence alignment and snippet summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 10–20, Online. Association for Computational Linguistics.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to summarize radiology findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213, Brussels, Belgium. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. DialogLM: Pretrained Model for Long Dialogue Understanding and

Summarization. *Computing Research Repository*, arXiv:2109.02492. Version 2.

Appendix

A Hyperparameters

The pre-training hyperparameters are listed in Table A.1 and the fine-tuning hyperparameters are listed in Table A.2. Each model was trained using a single NVIDIA V100 GPU. Mixed precision training and gradient checkpointing were used as needed in order to fit the larger models into memory.

B NER Model Performance

We fine-tune RoBERTa (pre-trained on MIMIC-III) on the i2b2 2010 dataset using the approach of Yang et al. (2020) in order to use it as a clinical concept extraction model for concept-based evaluation. We show our performance on the fine-tuning dataset in Table B.1 and compare it to theirs. While we were not able to fully match their results, we believe this is due to the fact that the i2b2 2010 dataset is no longer available in its original form. Nonetheless, we also achieve strong results that are suitable for our purposes.

Model	*P*	*R*	*F1*
Yang et al. (2020)	89.63	90.26	89.94
Ours	87.80	88.58	88.19

Table B.1: Comparison of clinical named entity recognition models.

C Additional Evaluation

The complete ROUGE evaluation results are shown in Table C.1, which shows precision and recall in addition to the F1 score. Similarly, Table C.2 shows precision and recall for the two concept-based evaluation methods.

Parameter	BART		LED		DialogLED	
	base	*large*	*base*	*large*	*base*	*large*
Maximum encoder length	1024	1024	5120	5120	5120	5120
Maximum decoder length	1024	1024	1024	1024	1024	1024
Text infilling ratio[a]	0.3	0.3	0.3	0.3	0.3	0.3
Window ratio	0.1	0.1	0.1	0.1	0.1	0.1
Maximum window size	512	512	512	512	512	512
Text infilling ratio[b]	0.15	0.15	0.15	0.15	0.15	0.15
Speaker mask ratio	0.5	0.5	0.5	0.5	0.5	0.5
Learning rate	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}
Batch size	8	8	8	8	8	8
Epochs	3	0.6	3 (1)[c]	0.4	3 (1)[c]	0.4 (0.2)[c]
Warm-up ratio	0.01	0.01	0.01	0.01	0.01	0.01
Weight decay	0.001	0.001	0.001	0.001	0.001	0.001
Maximum gradient norm	1.0	1.0	1.0	1.0	1.0	1.0

Table A.1: Hyperparameters used for continued pre-training. We differentiate between BART-style noise, which uses text infilling across the entire input (*a*), and window-based denoising, which only performs text infilling within the window (*b*) and masks speakers separately. Both types of denoising are investigated for LED and DialogLED. LED and DialogLED sometimes use different number of epochs during training for BART-style and window-based denoising (*c*). No sentence or turn permutation is used.

Parameter	BART		LED		DialogLED	
	base	*large*	*base*	*large*	*base*	*large*
Maximum encoder length	1024	1024	5120	5120	5120	5120
Maximum decoder length	1024	1024	1024	1024	1024	1024
Learning rate	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}
Batch size	8	8	8	8	8	8
Maximum epochs	30	30	30	30	30	30
Warm-up steps	200	200	200	200	200	200
Weight decay	0.001	0.001	0.001	0.001	0.001	0.001
Maximum gradient norm	0.1	0.1	0.1	0.1	0.1	0.1
Steps between evaluation	50	50	50	50	50	50
Early-stopping patience	5	5	5	5	5	5
Number of beams	5	5	5	5	5	5
Maximum generation length	512	512	512	512	512	512
No repeat n-gram size	3	3	3	3	3	3

Table A.2: Hyperparameters used for fine-tuning.

		ROUGE-1			ROUGE-2			ROUGE-L		
Model	**Pre-train**	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
BART-base	—	**44.71**	34.73	35.19	**17.05**	13.29	13.28	**31.73**	25.81	25.43
BART-base	BART	43.12	38.44	36.83	16.71	14.87	**14.13**	30.51	28.43	26.53
LED-base	—	43.26	37.03	36.01	16.36	14.03	13.49	30.65	27.46	25.99
LED-base	window	42.68	38.97	**36.96**	16.39	15.11	14.12	30.27	28.95	**26.72**
LED-base	BART	42.29	38.65	36.65	15.82	14.44	13.60	29.85	28.50	26.31
DialogLED-base	—	39.09	40.39	36.07	14.25	14.94	13.14	26.67	29.08	25.13
DialogLED-base	window	38.98	**42.04**	36.85	14.64	**15.89**	13.79	26.62	**30.45**	25.74
DialogLED-base	BART	39.56	41.49	36.79	14.67	15.53	13.59	27.24	30.14	25.88
BART-large	—	38.30	46.09	38.25	14.86	18.01	14.78	26.24	33.17	26.65
BART-large	BART	**42.73**	41.80	38.47	**16.64**	16.27	14.89	**29.83**	30.60	**27.37**
LED-large	—	37.00	46.11	37.29	13.70	17.44	13.83	25.43	33.40	26.09
LED-large	window	40.50	44.62	**38.88**	15.69	17.35	**14.96**	27.83	32.16	27.19
LED-large	BART	38.28	46.05	38.07	14.69	17.81	14.56	26.56	33.41	26.82
DialogLED-large	—	33.83	49.98	37.04	12.53	18.81	13.74	22.83	35.95	25.55
DialogLED-large	window	34.06	50.27	37.26	12.89	19.37	14.15	23.01	36.20	25.73
DialogLED-large	BART	33.34	**53.29**	37.73	12.88	**20.72**	14.56	22.25	**37.79**	25.69

Table C.1: Complete ROUGE evaluation results on the summarization test set.

		UMLS			NER		
Model	**Pre-train**	*P*	*R*	*F1*	*P*	*R*	*F1*
BART-base	—	56.65	18.28	24.27	**76.52**	26.66	36.22
BART-base	BART	**57.05**	22.93	28.63	70.93	32.56	40.24
LED-base	—	52.77	21.21	26.40	69.62	31.61	38.59
LED-base	window	53.04	21.88	27.45	70.29	32.43	39.96
LED-base	BART	53.29	22.72	28.47	66.81	34.98	41.39
DialogLED-base	—	51.99	26.63	31.22	66.34	**38.02**	42.66
DialogLED-base	window	53.51	26.35	31.62	65.27	36.31	41.87
DialogLED-base	BART	55.51	**28.36**	**33.33**	69.48	36.64	**42.72**
BART-large	—	**53.95**	30.52	35.19	66.39	42.92	47.52
BART-large	BART	46.97	23.49	27.77	**66.84**	37.42	43.45
LED-large	—	43.80	28.01	30.45	51.84	44.27	43.97
LED-large	window	51.15	27.13	32.03	65.93	41.39	46.78
LED-large	BART	52.93	31.51	35.33	63.92	43.54	47.15
DialogLED-large	—	44.14	30.43	32.36	54.46	48.84	47.23
DialogLED-large	window	46.29	32.22	34.05	53.48	46.98	45.86
DialogLED-large	BART	49.60	**37.88**	**38.90**	61.70	**52.18**	**51.57**

Table C.2: Complete concept-based evaluation results on the summarization test set.