LongBoX: Evaluating Transformers on Long-Sequence Clinical Tasks

Anonymous ACL submission

Abstract

Many large language models (LLMs) for medicine have largely been evaluated on short texts, and their ability to handle longer sequences such as a complete electronic health record (EHR) has not been systematically explored. Assessing these models on long sequences is crucial since prior work in the general domain has demonstrated performance degradation of LLMs on longer texts. Motivated by this, we introduce LONGBOX, a collection of seven medical datasets in text-to-text format, designed to investigate model performance on long sequences. Preliminary experiments reveal that both medical LLMs (e.g., BioGPT) and strong general domain LLMs (e.g., FLAN-T5) struggle on this benchmark. We further evaluate two techniques designed for long-sequence handling: (i) local-global attention, and (ii) Fusion-in-Decoder (FiD). Our results demonstrate mixed results with long-sequence handling - while scores on some datasets increase, there is substantial room for improvement. We hope that LONGBOX facilitates the development of more effective long-sequence techniques for the medical domain[1].

1 Introduction

In recent years, the exponential increase in machine-readable text in the medical domain such as electronic health records (EHRs) has sparked a growing interest in the development of pretrained medical language models (Lewis et al., 2020). Over the years, many large language models (LLMs) have been developed in these domains such as BioGPT (Luo et al., 2022), BioMedLM (Venigalla et al., 2022), GatorTRONGPT (Peng et al., 2023) and MedPaLM (Singhal et al., 2022). These LLMs have been evaluated on a wide range of medical tasks, but most tasks have only involved short texts. Many real-world medical tasks on the other hand require models to make predictions from longer texts, such as a summary from a patient visit or a series of EHRs for a patient, hence evaluating performance on longer texts is crucial. While this problem has been tackled in the general domain (Shaham et al., 2022; Tay et al., 2021), model ability to handle long sequences in the clinical domain is under-explored (More related work in App. A).

To tackle this, we propose LONGBOX, a collection of seven carefully-curated clinical datasets, which can measure performance of models on long sequences, converted to a unified text-to-text format. LONGBOX incorporates three task types: text classification, relation extraction and multi-label classification, and several types of clinical inputs such as discharge summaries and longitudinal records. Most importantly, for all datasets, input texts typically contains thousands of words.

We first benchmark the performance of widely used high-performing LLMs on LONGBOX from general domain: LLaMA-2 (Touvron et al., 2023), GPT-Neo (Black et al., 2022), FLAN-T5 (Chung et al., 2022) and from medical domain: SciFive (Phan et al., 2021), In-BoXBART (Parmar et al., 2022), Clinical-T5 (Lu et al., 2022), BioGPT (Luo et al., 2022), and BioMedLM (Venigalla et al., 2022). Our results reveal that these models struggle on all datasets from LONGBOX achieving an average score of $\sim 52\%$. Next we evaluate two long sequence techniques that have shown promise in the general domain: (i) local-global attention (e.g., LongT5 (Guo et al., 2022)), and (ii) Fusion-in-Decoder (FiD) (Izacard and Grave, 2021) (w/ SciFive and Clinical-T5). These methods achieve mixed results on LONGBOX, further highlighting the need for our benchmark. We further evaluate two long sequence clinical models, i.e., Clinical-Longformer and Clinical-BigBird (results are discussed in App. D). We hope LONGBOX facilitates the development of better long sequence handling techniques for medical text.

[1]Data and source code are available at <anonymous link>

Dataset	Document Types	# of Samples			Avg. Tokens	Max. Tokens
		Train	Val	Test		
Smoking 2006	DS	358	40	104	1251.98	3858
Obesity 2008	DS	11552	805	7239	1920.47	4494
Assertions 2010	DS, PR	7073	1259	11013	2237.40	5805
Temporal RE	DS	31513	2554	22643	1245.68	2866
RFHD 2014	LR	4243	280	2516	1194.14	4660
Cohort Selection	LR	2626	1118	1118	6970.14	25637
ADE 2018	DS	36348	2346	20593	4356.43	11632

Table 1: An overview of document types used to create the dataset, along with a statistical analysis of each dataset. DS: Discharge Summaries, PR: Progress Reports, LR: Longitudinal Records

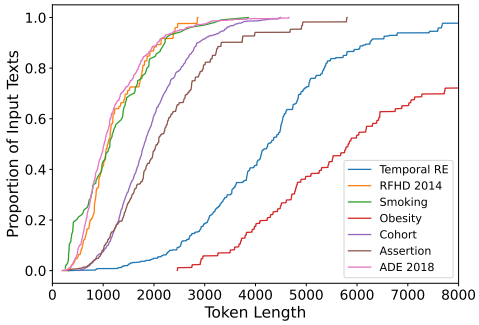

Figure 1: Cumulative distributions of input token lengths for all LONGBOX datasets.

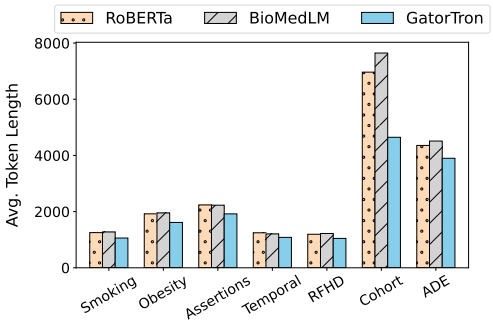

Figure 2: Average token length comparison between GatorTron and RoBERTa for all LONGBOX datasets.

2 LongBoX

LONGBOX contains seven clinical datasets curated from *n2c2 NLP Research* collection[2]:(1) Smoking Challenge 2006 (Uzuner et al., 2008), (2) Obesity Challenge 2008 (Uzuner, 2009), (3) Assertions Challenge 2010 (Uzuner et al., 2011), (4) Temporal Relations 2012 (Sun et al., 2013), (5) Heart Disease 2014 (Kumar et al., 2015), (6) Cohort Selection 2018 (Stubbs et al., 2019), and (7) Adverse Drug Events (ADE) 2018 (Henry et al., 2020). Table 1 presents the type of input text, dataset splits, and token length statistics for each dataset, with further details in Appendix B.

2.1 Qualitative Analysis

Length Analysis: Table 1 presents the average and maximum input token lengths of test sets per dataset after tokenization with the RoBERTa-large tokenizer, which range from 1194-6970 and 2866-25637 respectively. Additionally, Figure 1 displays cumulative distributions of input token lengths for each dataset (cut off at $8k$ for visibility) - given a token length x, the Y-axis indicates the proportion of inputs in the test set with token length $\leq x$. Despite considerable variation across datasets, the maximum token limit for most LLMs (1024) is within 40^{th} percentile range for most datasets, and most of instances in each dataset exceed $3k$ tokens.

Comparing input lengths under domain-specific tokenizers (RoBERTa, BioMedLM, and GatorTron): To assess whether clinical tokenizers significantly reduce text lengths over general domain tokenizers, we compare text lengths post tokenization by GatorTron (clinically-tailored), BioMedLM (biomedically-tailored) and RoBERTa (general domain). We tokenize test sets of all datasets using these three tokenizers. Figure 2 presents average token lengths for the test set of each dataset (on the X-axis) from LONGBOX. It is evident that the clinical tokenizer generates shorter token lengths compared to the biomedical and general domain tokenizers, though differences are often small. We also observe that average token lengths for biomedical *vs.* general tokenizer are nearly similar. Notably, difference between average token lengths for clinical *vs.* biomedical or general tokenizers becomes larger as input length increases, particularly observed in cohort selection.

[2] https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

Dataset	*Enc.* + *Dec.* Models				*Dec.* Models			
	FLAN-T5	**In-BoXBART**	**SciFive**	**Clinical-T5**	**GPT-Neo**	**BioGPT**	**BioMedLM**	**LLaMA-2**
Smoking 2006	55.77	58.65	60.58	**64.41**	3.85	**56.73**	2.89	38.46
Obesity 2008	68.28	**71.86**	**71.86**	70.55	51.50	33.21	71.86	**84.78**
Assertions 2010	**68.86**	67.95	67.83	68.17	61.07	63.77	67.83	**73.09**
Temporal RE	**56.53**	56.36	56.03	56.27	38.10	10.75	37.29	**54.22**
RFHD 2014	64.99	**66.64**	64.58	65.57	58.59	11.34	44.34	**74.76**
Cohort Selection	45.53	**47.67**	41.05	47.41	51.23	53.43	**58.95**	47.41
ADE 2018	19.07	17.62	**19.22**	18.56	9.70	4.96	8.79	**23.97**

Table 2: Performance of *Enc.* + *Dec.* and *Dec.* models on LONGBOX. All results are presented in %.

3 Experiments and Results

3.1 Experimental Setup

Models: We benchmark eight models from two architecture families: (i) four Encoder (*Enc.*) + Decoder (*Dec.*) models (FLAN-T5-Large from general domain; SciFive-Large, In-BoxBART, and Clinical-T5-Large from medical domain), and (ii) four Decoder (*Dec.*) only models (LLaMA-2-7B and GPT-Neo-1.3B from general domain; BioGPT-1.5B and BioMedLM-2.7B from medical domain). In addition, we evaluate two long sequence models. The first one is LongT5-Large, which enables a T5 encoder (Raffel et al., 2020) to more efficiently handle long sequences by leveraging local-global attention sparsity patterns. The second is Fusion-in-Decoder (FiD), which breaks each input into smaller chunks, encodes them using an encoder-decoder model and then fuses encoded chunks in the decoder while generating output. We experiment with both SciFive and Clinical-T5 as the base encoder-decoder models for FiD.

Experimental Details: For all the models, we reframed all the datasets as text generation tasks and provide every (input, output) pair in text-to-text format. However, when training *Dec.* models using this setting (except for LLaMA-2), we observe poor performance in majority cases on classification and relation extraction - they either produce malformed labels or just generate continuations for the input text instead of generating the output label. While we did not investigate this deeper, this indicates that long inputs might be particularly problematic for *Dec.* models. Based on these observations, we further investigate a different setup for *Dec.* models (except LLaMA-2 since it achieves good performance in above setting) on these tasks: the final prediction is made by first encoding the input, then applying a classification head to the last token. Results are presented in Appendix C. All *Enc.*+*Dec.* and *Dec.* models have an input length of 1024 tokens, while LongT5 and FiD are evaluated with three different token lengths: 2048, 3072, and 4096. More details are presented in App. E.

Metrics: For all classification and relation extraction tasks in LONGBOX, we report performance using the *Accuracy* metric. However, for RFHD 2014, which is the only dataset for multi-label classification, we use the F_1-score metric. For Dec. models, we report *lenient accuracy* for all tasks, which post-processes predictions to exclude any unnecessary text generated aside from the predicted label to determine the final *accuracy*.

3.2 Results

Table 2 presents the performance of all general and medical domain LLMs (baseline models) benchmarked on LONGBOX, while Tables 3 and 4 show the performance of the two long sequence techniques we test.

Baseline Models: Table 2 shows that overall, average performance of all benchmarked models on LONGBOX is low ($\sim 52\%$). Among *Enc.* + *Dec.* models, medical LLMs generally outperform general domain models on most datasets (five of seven), and are competitive with each other. For *Dec.* only models, we see the reverse - LLaMA-2 outperforms medical LLMs on most datasets (five of seven). We also observe that all models consistently exhibit lower scores on datasets with higher input lengths such as ADE 2018 and Cohort Selection (see Table 1 for lengths), indicating that long input techniques could help. Lastly, as model size increases, we see that capability to handle longer texts improves; for instance LLaMA-2 (7B) improves results on five of seven datasets, compared to other models (<2.7B).

Long Sequence Techniques: From Table 3 and Table 4, we see that adding more input context provides mixed results - only improving performance over baseline models on some datasets. We also observe that performance on many datasets continues to improve with increasing input length (from 2048 to 4096 tokens). We further qualitatively analyze the mixed performance of long sequence

Dataset	FiD (w/SciFive)			FiD (w/ClinicalT5)		
	2048	3072	4096	2048	3072	4096
Smoking 2006	60.26 0.32% ↓	62.03 1.45% ↑	64.42 3.84% ↑	56.73 3.53% ↓	60.58 1.45% ↓	60.58 3.84% ↓
Obesity 2008	64.82 7.04% ↓	71.32 0.54% ↓	73.15 1.29% ↑	64.36 0.46% ↓	73.00 1.68% ↑	74.20 1.05% ↑
Assertions 2010	67.14 0.69% ↓	66.95 0.88% ↓	66.71 1.12% ↓	66.71 0.43% ↓	66.92 0.03% ↓	67.06 0.35% ↑
Temporal RE	58.81 2.78% ↑	60.17 4.14% ↑	63.21 7.18% ↑	58.37 0.44% ↓	60.53 0.36% ↑	63.79 0.58% ↑
RFHD 2014	70.65 5.98% ↑	76.16 11.6% ↑	78.60 14.0% ↑	60.65 10.0% ↓	65.46 10.7% ↓	68.76 9.84% ↓
Cohort Selection	48.66 7.61% ↑	46.87 5.82% ↑	44.28 3.23% ↑	48.12 0.54% ↓	46.51 0.36% ↓	46.33 2.23% ↑
ADE 2018	17.58 1.64% ↓	29.15 9.93% ↑	46.94 27.7% ↑	17.73 0.15% ↑	29.36 0.21% ↓	47.07 0.13% ↑

Table 3: Performance of long document techniques, FiD (w/SciFive) and FiD (w/ClinicalT5), on LONGBOX. All results are presented in %. Green indicates improvement and red indicates degradation in performance comparison between the SciFive *vs.* FiD (w/SciFive) and FiD (w/ClinicalT5) *vs.* FiD (w/SciFive).

Dataset	LongT5		
	2048	3072	4096
Smoking 2006	53.85 10.6% ↓	58.65 5.76% ↓	55.77 8.64% ↓
Obesity 2008	71.87 0.01% ↑	76.79 4.93% ↑	77.73 5.87% ↑
Assertions 2010	67.85 1.01% ↓	68.07 0.79% ↓	67.76 1.10% ↓
Temporal RE	60.73 4.20% ↑	57.96 1.43% ↑	72.89 15.4% ↑
RFHD 2014	45.07 21.6% ↓	45.32 21.3% ↓	44.44 22.2% ↓
Cohort Selection	56.35 0.54% ↑	57.70 10.1% ↑	48.30 0.63% ↑
ADE 2018	18.12 1.10% ↓	17.83 1.39% ↓	46.58 27.4% ↑

Table 4: Performance of long document technique LongT5 on LONGBOX. All results are presented in %. Green indicates improvement and red indicates degradation in performance compared to the best performing $Enc. + Dec.$ model from Table 2.

techniques.

Clinical vs Biomedical Base Models for FiD: Table 3 shows that clinical pretraining shows marginal improvements over using FiD with a biomedical base model on some datasets with the largest input token length (4096). We also observe that FiD (w/ClinicalT5) shows mixed performance on many datasets for 2048 and 3072 input token lengths.

3.3 Qualitative Analysis

Why is LONGBOX difficult for long document models? We first perform a qualitative error analysis on one dataset on which long input techniques provide no improvement over baselines: cohort selection. We randomly sample 50 cases which both techniques get wrong and observe three categories of errors. The first one is caused due to very few and/or late occurrences (i.e., outside our maximum length of 4096 tokens) of informative cues needed for the task. The second one stems from a lack of awareness of EHR document structure (e.g., family history does not contain conditions present in current patient) or ability to deal with longitudinal records (e.g., later test results override earlier ones). The third category is not caused due to input length, but rather consists of errors caused by the presence of comorbidities, similar symptoms, etc., which require precise clinical inference.

Why do models lag behind human performance? Despite mixed results, long document techniques do improve perfomance on some datasets, but lag behind human performance. We analyze 50 randomly sampled error cases from one dataset (obesity 2008) and observe the same three error categories, with ∼80% errors falling into the third category (requiring precise clinical inference). This indicate two potential avenues to push performance on LONGBOX: (i) exploring *relevant sentence selection* in addition to increased context length, and (ii) developing pretraining/finetuning techniques to equip models with the ability to handle document structure and longitudinality.

4 Conclusions

We introduced LONGBOX, a collection of seven carefully curated clinical datasets, aimed to comprehensively and systematically investigate performance of clinical LMs on long texts. LONGBOX covers three task types: text classification, relation extraction and multi-label classification and various input types like longitudinal records and discharge summaries. We benchmark the performance of eight general and medical domains LLMs on LONGBOX, and show that they do not achieve good performance. We also investigate two long sequence techniques and our results reveal that though these methods provide some benefit, there is substantial room for improvement. We believe that LONGBOX can serve as an important benchmark for developing long sequence techniques tailored to the clinical domain.

Limitations

Currently, LONGBOX is limited in terms of task variety since it primarily consists of different types of classification tasks. This is largely because it is challenging to find shareable datasets across various task types in the clinical domain, but we plan to further increase task variety in this benchmark. Additionally, we hope to expand our analysis to include the most recent large language models such as GPT-4 and ChatGPT on LONGBOX. Our observation that existing long document models still struggle on LONGBOX, also suggests that it may be interesting to conduct detailed analysis of different aspects such as model understanding of clinical document structure and better clinical tokenization, which we have left to future work.

References

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Samuel Cahyawijaya, Bryan Wilie, Holy Lovenia, Huan Zhong, MingQian Zhong, Yuk-Yu Nancy Ip, and Pascale Fung. 2022. How long is enough? exploring the optimal intervals of long-range clinical note language modeling. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 160–172, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Zican Dong, Tianyi Tang, Lunyi Li, and Wayne Xin Zhao. 2023. A survey on long text modeling with transformers. *arXiv preprint arXiv:2302.14502*.

Quentin Fournier, Gaétan Marceau Caron, and Daniel Aloise. 2021. A practical survey on faster and lighter transformers. *ACM Computing Surveys*.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Amy JH Kind and Maureen A Smith. 2008. Documentation of mandated discharge summary components in transitions from acute to subacute care. *Advances in patient safety: new directions and alternative approaches (Vol. 2: culture and redesign)*.

Vishesh Kumar, Amber Stubbs, Stanley Shaw, and Özlem Uzuner. 2015. Creation of a new longitudinal corpus of clinical narratives. *Journal of biomedical informatics*, 58:S6–S10.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*.

Qiuhao Lu, Dejing Dou, and Thien Nguyen. 2022. ClinicalT5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Mihir Parmar, Swaroop Mishra, Mirali Purohit, Man Luo, Murad Mohammad, and Chitta Baral. 2022. In-BoXBART: Get instructions into biomedical multi-task learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 112–128, Seattle, United States. Association for Computational Linguistics.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al.

2023. A study of generative large language model for medical research and healthcare. *arXiv preprint arXiv:2305.13523*.

Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. 2022. SCROLLS: Standardized CompaRison over long language sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 12007–12021, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuqi Si and Kirk Roberts. 2021. Three-level hierarchical transformer networks for long-sequence and multiple clinical documents classification. *arXiv preprint arXiv:2104.08444*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Amber Stubbs, Michele Filannino, Ergin Soysal, Samuel Henry, and Özlem Uzuner. 2019. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, 26(11):1163–1171.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena: A benchmark for efficient transformers. In *International Conference on Learning Representations*.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Özlem Uzuner. 2009. Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(4):561–570.

Özlem Uzuner, Ira Goldstein, Yuan Luo, and Isaac Kohane. 2008. Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1):14–24.

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*, 18(5):552–556.

A Venigalla, J Frankle, and M Carbin. 2022. Biomedlm: a domain-specific large language model for biomedical text. *MosaicML. Accessed: Dec*, 23:3.

A Related Work

Prior work in the general domain has developed benchmarks to evaluate the ability of transformer-based models to handle long sequence tasks (Tay et al., 2021; Shaham et al., 2022). These benchmarks motivated the design of several techniques capable of handling long input sequences (see Dong et al. (2023); Tay et al. (2022); Fournier et al. (2021) for detailed surveys), which can broadly be divided into two categories: (i) architecture-focused approaches (e.g., developing sparse or hierarchical attention mechanisms), and (ii) data-focused approaches (e.g., chunking or subselecting input). However, most of these methods have not been systematically and broadly tested in the clinical domain due to the lack of a comprehensive benchmark which we try to address.

In the clinical domain, some prior work has explored architecture-focused long document approaches (Si and Roberts, 2021; Li et al., 2022; Cahyawijaya et al., 2022), however, their evaluation is limited to a handful of tasks. LONGBOX, on the other hand, covers a broad range of tasks and datasets in the clinical domain with longer input token lengths ($> 5k$ in many cases) for more comprehensive and systematic evaluation.

B Benchmark Details

We provide a comprehensive overview of all datasets in LONGBOX, along with descriptions of diverse document types that were annotated to create these datasets.

B.0.1 Document Types

Discharge Summaries are clinical notes containing details about why a person was admitted, diagnosis, medical regimen and response to their diagnosis, medical condition at discharge time, and after discharge care such as medications to continue at home (Kind and Smith, 2008). These summaries are in long text format but often not organized.

Progress Reports are clinical documents that form the basis of the next plan of treatment. They consist of assessment, diagnosis, planning, intervention, and evaluation sections.

Longitudinal Records are clinical documents that aggregate information from various sources in the health care system.

B.0.2 Dataset Overview

Smoking 2006 (Uzuner et al., 2008): Given discharge summaries for patients, the task is to categorize the smoking status of a patient into: (1) Past Smoker, (2) Current Smoker, (3) Smoker, (4) Non-Smoker, and (5) Unknown. This dataset was released as part of the n2c2 challenge in 2006.

Obesity 2008 (Uzuner, 2009): Based on discharge summaries, the task is to determine the presence of 15 different diseases such as asthma, and diabetes, which are potential indicators of obesity. The goal here is to categorize the presence of disease into: (1) Present, (2) Absent, (3) Questionable, and (4) Unmentioned. This dataset was released as part of the n2c2 challenge in 2008.

Assertions 2010 (Uzuner et al., 2011): Given discharge summaries as well as progress reports of patients, the task is to classify the occurrence of a concept into 6 categories: (1) Present, (2) Absent, (3) Hypothetical, (4) Possible, (5) Associated with someone else, and (6) Conditional. The concept can be medical problems, treatments, and tests. This dataset was released as part of the n2c2 challenge in 2010.

Temporal Relations 2012 (Sun et al., 2013): The dataset consists of discharge summaries. Given a clinically significant event and time entity, the task is to find the type of relationship between them - BEFORE (event happens before given temporal expression), AFTER (event happens after given temporal expression), SIMULTANEOUS (event happens on given temporal expression), OVERLAP (event overlaps with temporal expression), BEGUN_BY (event started on given temporal expression), ENDED_BY (event ended on given temporal expression), DURING (event happens during given temporal expression), and BEFORE_OVERLAP (event started before and lasts during given temporal expression). This dataset was released as part of the n2c2 challenge in 2012.

Heart Disease 2014 (Kumar et al., 2015): This dataset consists of longitudinal medical records. The task here is to find indicators of a given condition in the text and classify them into "Present" and "Not present". For instance, the indicator for Diabetes can be different aspects such as the patient mentioning having diabetes, high glucose, and high HBA1c levels. This dataset was released as part of the n2c2 challenge in 2014.

Dataset	RoBERTa	Clinical-Longformer			Clinical-BigBird		
	512	2048	3072	4096	2048	3072	4096
Smoking 2006	61.54	63.46	56.73	59.62	78.85	80.77	82.69
Obesity 2008	71.86*	71.86*	71.86*	71.86*	71.86*	71.86*	71.86*
Assertions 2010	67.84*	72.52	67.84*	75.00	67.84*	67.84*	77.46
Temporal RE	54.01*	54.01*	54.01*	54.01*	55.75	54.01*	58.3
Cohort Selection	58.95*	58.95*	56.08	58.95*	58.95*	58.95*	57.87
ADE 2018	17.65*	17.65*	17.65*	18.01	17.65*	17.65*	17.65*

Table 5: Performance of RoBERTa-large, Clinical-Longformer, and Clinical-BigBird models on LONGBOX. All results are presented in %. * denotes that the model has only generated labels corresponding to the majority classes.

Cohort Selection (Stubbs et al., 2019): In this dataset, the goal is to classify whether a patient meets or does not meet specific criteria for participation in clinical trials. Clinical trials have certain criteria for including a patient in the trial group. The dataset includes 13 defined criteria such as MAJOR-DIABETES (Major diabetes-related complication), ALCOHOL-ABUSE (Current alcohol use over weekly recommended limits), and ENGLISH (Patient must speak English). This dataset was released as part of the n2c2 challenge in 2018.

ADE 2018 (Henry et al., 2020): Given discharge summaries, the task here is to classify the relationship between a drug and another related entity such as Strength-Drug (e.g., 20mg), Dosage-Drug (e.g., 1 tab per day), Duration-Drug (e.g., 5-day course), Frequency-Drug (e.g., every 4-6 hrs), Form-Drug (e.g., tablet, capsule), Route-Drug (e.g., intraperitoneal, IM), Reason-Drug (reason/disease for which the medication is prescribed), and ADE-Drug (side effect caused by the drug). This was another dataset released as part of the n2c2 challenge in 2018.

C Additional Results - Dec. Models

In this section, we provide results for our further investigation on a different setup for *Dec.* models on LONGBOX: the final prediction is made by first encoding the input, then applying a classification head to the last token. Results are presented in Table 6. From the Table 6, it is evident that applying a classification head to the last token improve the model performances in majority tasks by large margin.

D Additional Results - Long Sequence Clinical Models

In this section, we present an evaluation of long sequence clinical models - Clinical LongFormer and Clinical BigBird (Li et al., 2022). Given that these models are based on the RoBERTa encoder-only architecture, we compare their performance against the RoBERTa-large model. It's well-known that smaller models are susceptible to class imbalance, and our findings reflect this trend: in the majority of cases, these models predominantly predict labels corresponding to the majority class only. Our evaluation specifically focuses on single-label tasks within the LONGBOX. As our primary objective is to assess these models on the LONGBOX and emphasize the necessity of our benchmark, we intend to conduct further experiments aimed at enhancing their performance in future.

Dataset	GPT-Neo	BioGPT	BioMedLM
Smoking 2006	58.65 54.8% ↑	59.62 2.89% ↑	50.58 47.69% ↑
Obesity 2008	73.08 21.58% ↑	71.86 38.65% ↑	71.75 0.11% ↓
Assertions 2010	70.87 9.8% ↑	67.63 3.86% ↑	66.01 1.82% ↓
Temporal RE	46.37 8.27% ↑	48.54 37.79% ↑	48.96 11.67% ↑
RFHD 2014	34.13 24.46% ↓	2.9 8.44% ↓	33.98 10.36% ↑
Cohort Selection	55.90 4.67% ↑	51.52 1.91% ↓	46.60 12.35% ↓
ADE 2018	21.95 12.25% ↑	17.46 12.5% ↑	17.29 8.5% ↑

Table 6: Comparison of different approach for *Dec.* models on LONGBOX w.r.t. Table 2. All results are presented in %.

E Additional Experimental Setup

For better comparability, we use the same hyperparameter settings for all runs: training is run for 3 epochs, with a batch size of 32 and an initial learning rate of 5e-5. The experiments were conducted on A6000 and A100 NVIDIA GPUs.