Supplementary: Building Patient Journeys in Hebrew: A Domain-Specific Language Model for Clinical Timeline Extraction

A Full Results

Table 1 summarizes the final results for all model variations across the three downstream tasks.

B Common Words

In Figure 1, we present the most common words in the clinical-notes corpus compiled for training our model.

C Common Added Tokens

In Figure 2, we present the most frequent tokens added to the tokenizer's vocabulary using the simple vocabulary adaptation approach. All tokens are in Hebrew, and most represent sub-words relevant to the medical domain (e.g., "in her exam", "neuro-", "antibioti-").

D Tokenizer Evaluation Metrics

Corpus Token Count (CTC). Given a vocabulary V with a corresponding tokenizer T_V and a document d, the tokenizer T_V segments d into a series of K_d tokens:

$$T_V(d) = t_1, \ldots, t_i, \ldots, t_{K_d} \tag{1}$$

with all $t_i \in V$. Given a corpus of documents D, CTC is defined as the total number of tokens used in each segmentation:

$$CTC(D, T_V) = \sum_{d \in D} K_d \tag{2}$$

Compression Rate (CR). Given a corpus D, we have:

$$compression\_rate(D, T_V) = \frac{CTC(D, T_V)}{word\_count(D)} \tag{3}$$

With $word\_count(D)$ being a function that calculates the number of white-space delimited words in all documents on the corpus D.

E Additional Epochs

All reported results are based on one epoch of continual pretraining. We extend this to two and three epochs to assess the impact of longer training. As shown in Figure 3, additional epochs slightly improve Med-TRC but have no effect on Onc-TRC.

F Dataset Breakdown

Table 2 presents the departmental breakdown of the EHRs extracted from the hospital system. The records encompass a diverse range of types, including follow-ups, descriptions, recommendations, and even technical surgery course details.

References

[Beltagy *et al.*, 2019] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Pre-training of deep bidirectional transformers for language understanding, 2019.

[Gu *et al.*, 2021] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, October 2021.

[Lee *et al.*, 2019] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019.

[Peng *et al.*, 2019] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors, *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy, August 2019. Association for Computational Linguistics.

Table 1: Results on the two TRC tasks (relaxed F1-score). The numbers are means over 15 different runs, with standard deviations provided in parentheses.

Model	**Med-TRC**	**Onc-TRC**
DictaBERT 2.0 (Baseline)	85.5 (1)	76.6 (1)
Continual Pre-training (no vocab. adaptation)	**86.4** (1)	78.6 (2)
Continual Pre-training for 2 Epochs	86.1 (1)	78.5 (1)
Continual Pre-training for 3 Epochs	**86.4** (1)	78.4 (1)
Continual Pre-training + AdaLM	85.5 (1)	**79.1** (1)
Continual Pre-training + Simple	86.0 (1)	78.3 (1)
Continual Pre-training + Simple+FLOTA	86.0 (1)	78.5 (1)
Continual Pre-training + Simple + De-ID	86.1 (1)	78.7 (1)
English Models Baseline		
BlueBERT [Peng *et al.*, 2019]	39.1 (16)	62.2 (12)
mBERT [Devlin *et al.*, 2018]	82.6 (1)	76.6 (1)
BioBERT [Lee *et al.*, 2019]	54.5 (11)	67.4 (6)
BioLinkBERT [Yasunaga *et al.*, 2022]	54.8 (2)	69.2 (2)
PubmedBERT [Gu *et al.*, 2021]	52.5 (6)	71.5 (2)
SciBERT [Beltagy *et al.*, 2019]	58.1 (1)	73.6 (1)
BERT [Devlin *et al.*, 2019]	71.2 (2)	74.1 (1)

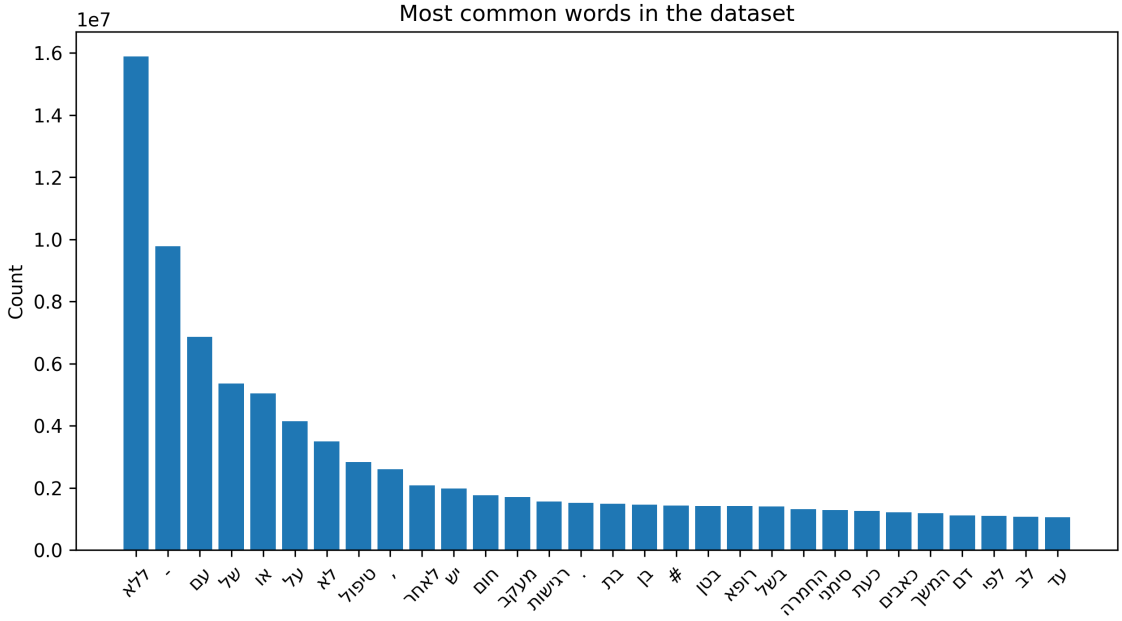

Figure 1: The most common words in our corpus

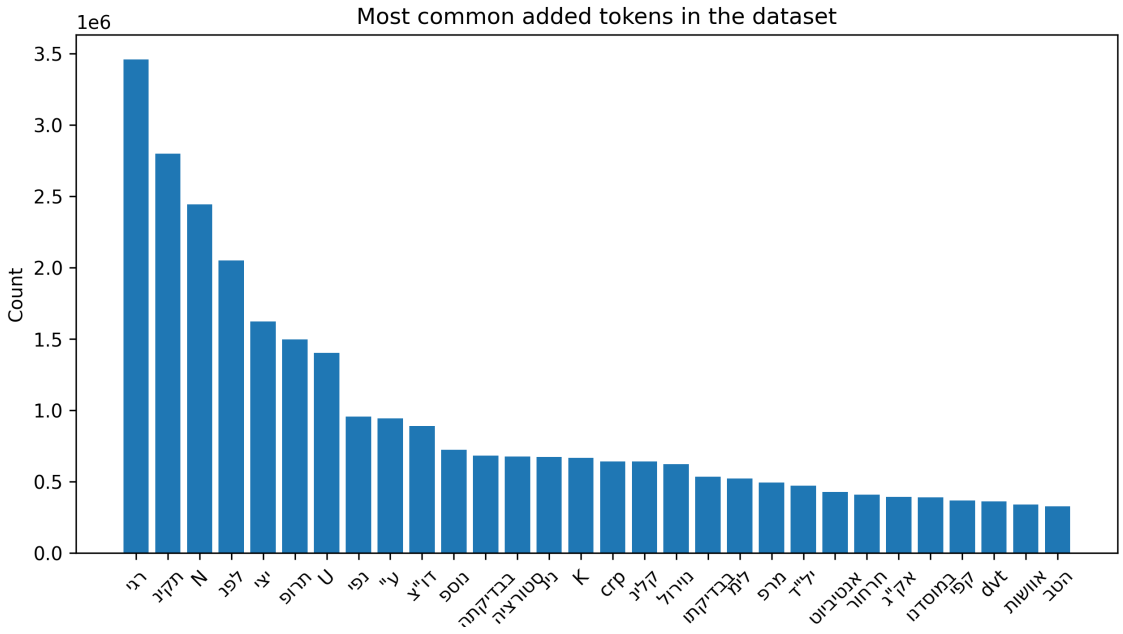

Figure 2: The most common tokens in the corpus that were added to the vocabulary using the simple vocabulary-adaptation method.

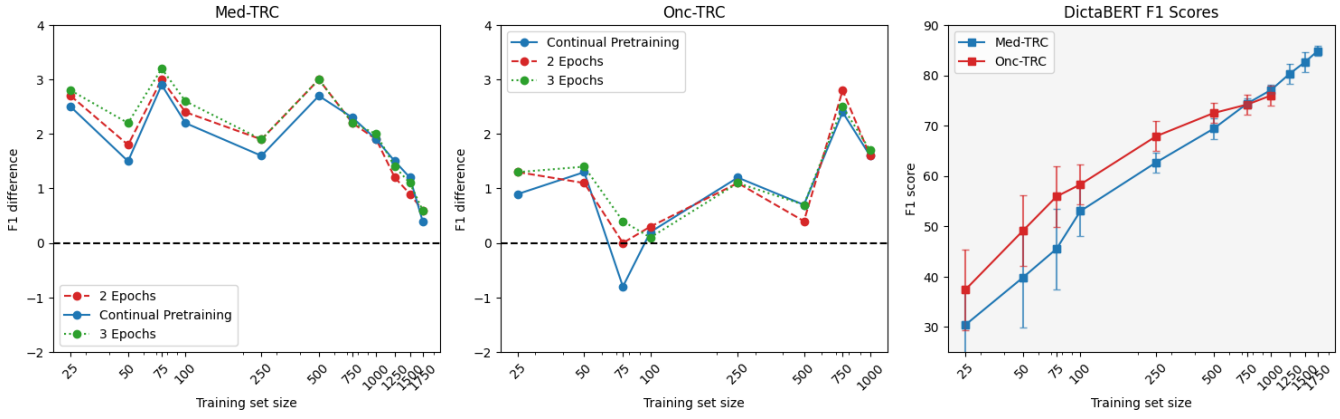

Figure 3: F1-score differences for various training set sizes. One, two or three pre-training epochs. The rightmost figure shows the absolute mean F1-score and the standard deviation of the **baseline** model.

Department	EHR Count
ER	2,385,411
Internal Medicine	1,325,387
Children ER	665,069
Labor and Delivery ER	530,601
Surgery Rooms	261,356
Surgery Rooms, Labor and Delivery	27,451
Ear, Nose, and Throat	16,148
Surgical Ward	4,312
Gynecology	3,481
Others	12,812
Total	**5,232,028**

Table 2: The distribution of EHRs by department, retrieved from the hospital for training our model.

[Yasunaga *et al.*, 2022] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links, 2022.