MedTok: Multimodal Medical Code Tokenizer

Xiaorui Su[1], Shvat Messica[1], Yepeng Huang[1], Ruth Johnson[1], Lukas Fesser[1], Shanghua Gao[1], Faryad Sahneh[2], Marinka Zitnik[1]

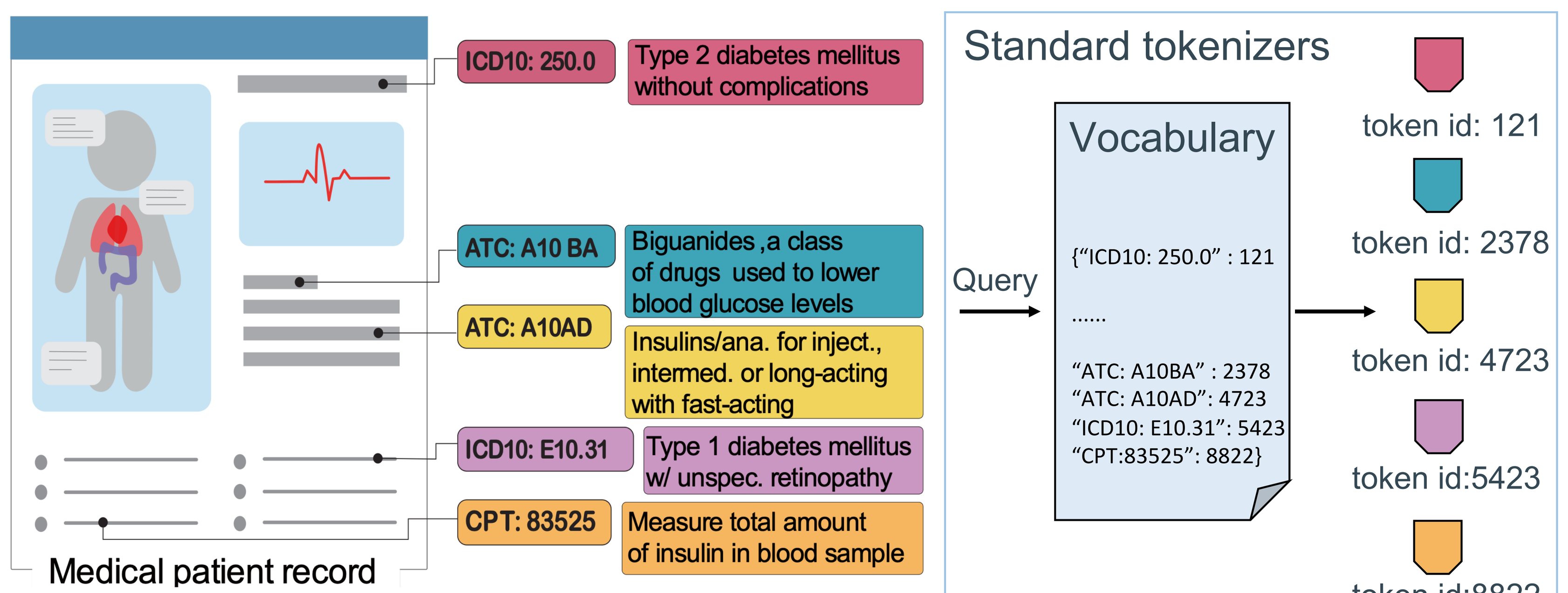
[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[2]Digital Data, Sanofi, Cambridge, MA, USA

xiaorui_su@hms.harvard.edu marinka@hms.harvard.edu

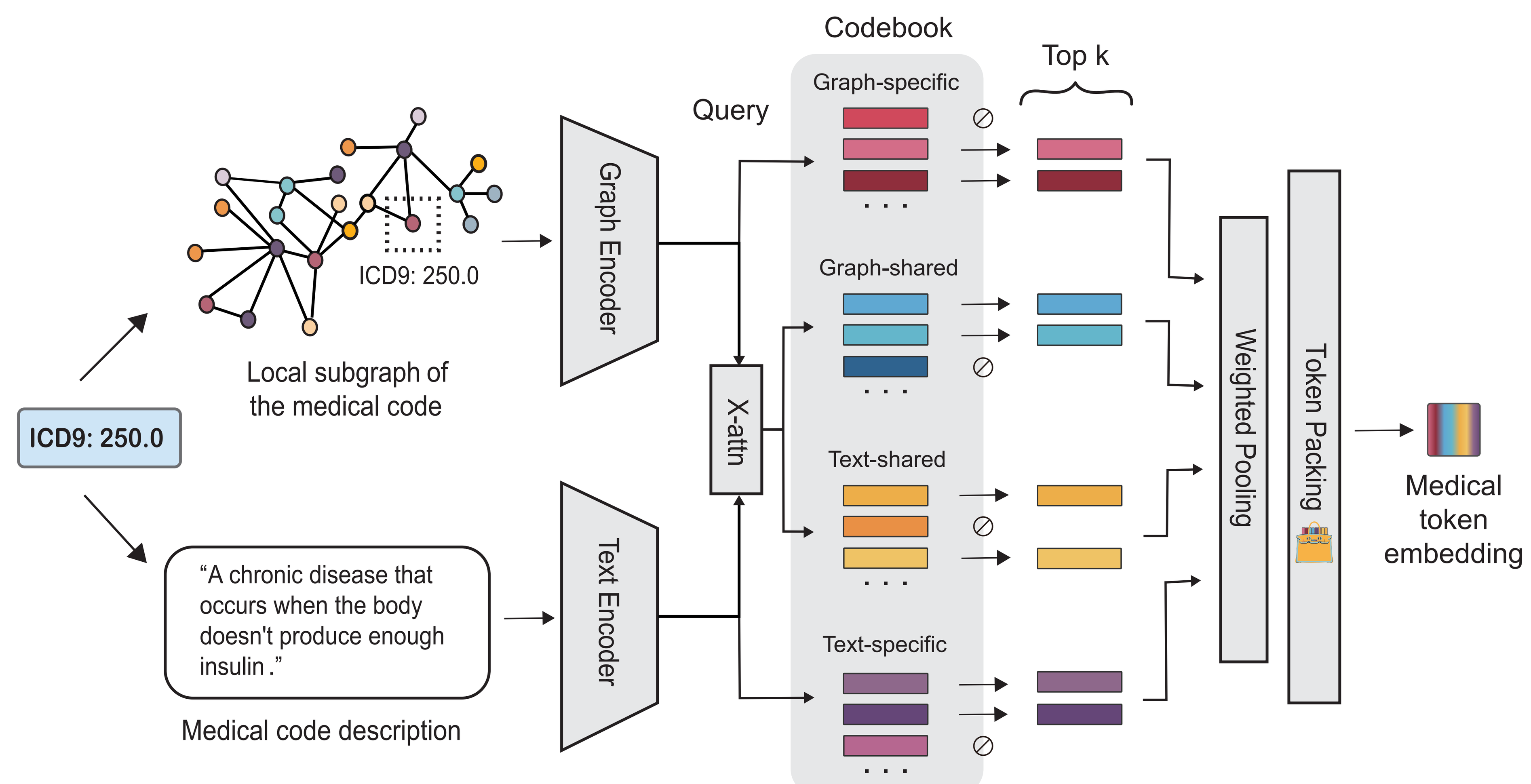
paper code

Standard tokenizers fail for medical codes



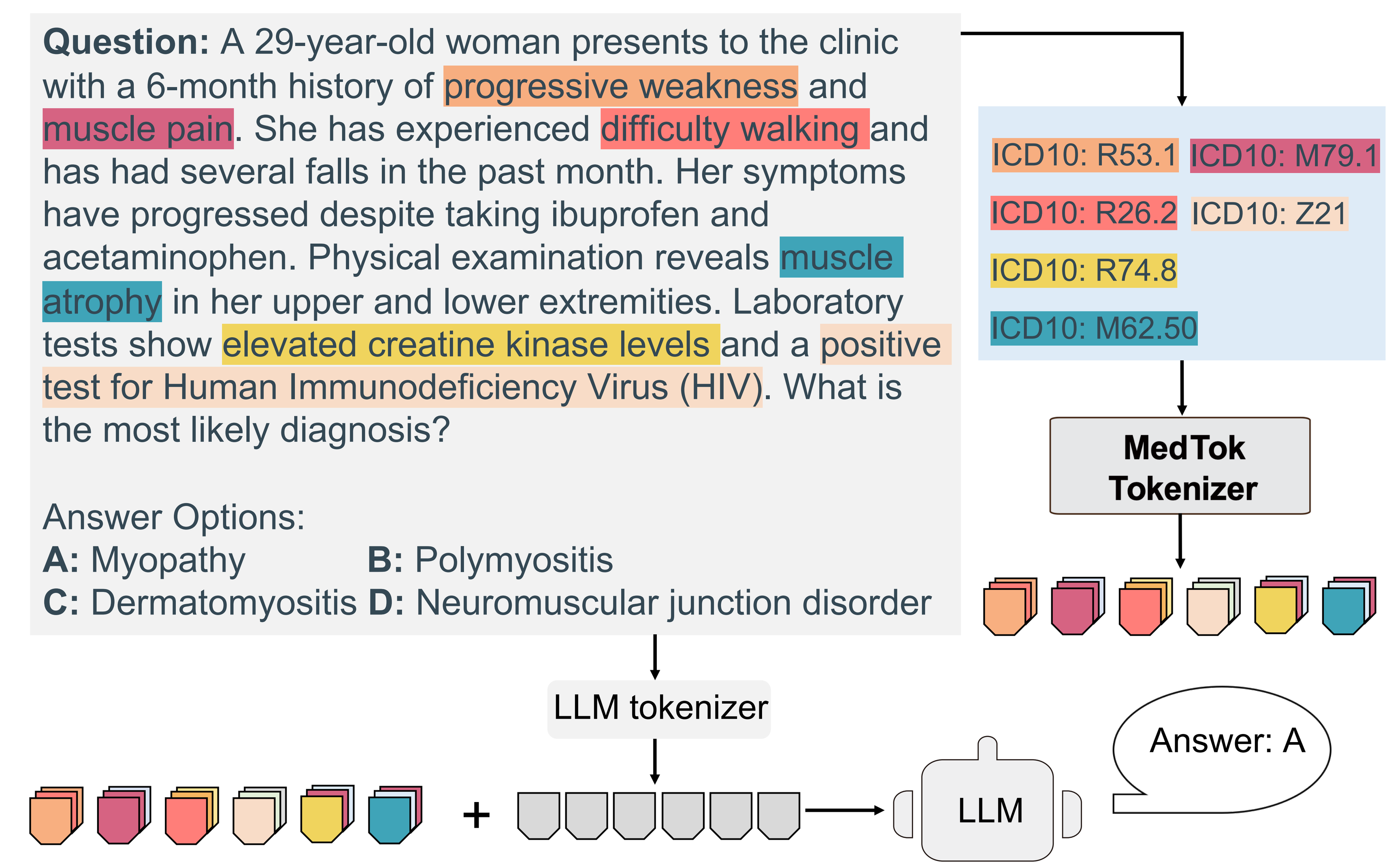
- Medical coding systems contain over 600,000 unique codes. Treating each code as a separate token leads to inefficient vocabulary expansion, increasing memory demands and fragmenting rare codes.
- Many coding systems encode structured dependencies, such as ATC code. Standard tokenizers, relying only on co-occurrence statistics, fail to capture hierarchical relationships, losing dependencies like disease co-occurrences and drug contraindications.
- Identical clinical concepts often appear under different codes across terminologies. Standard tokenization treats them as separate tokens, creating redundancy and complicating cross-system data integration.

Overview of MedTok

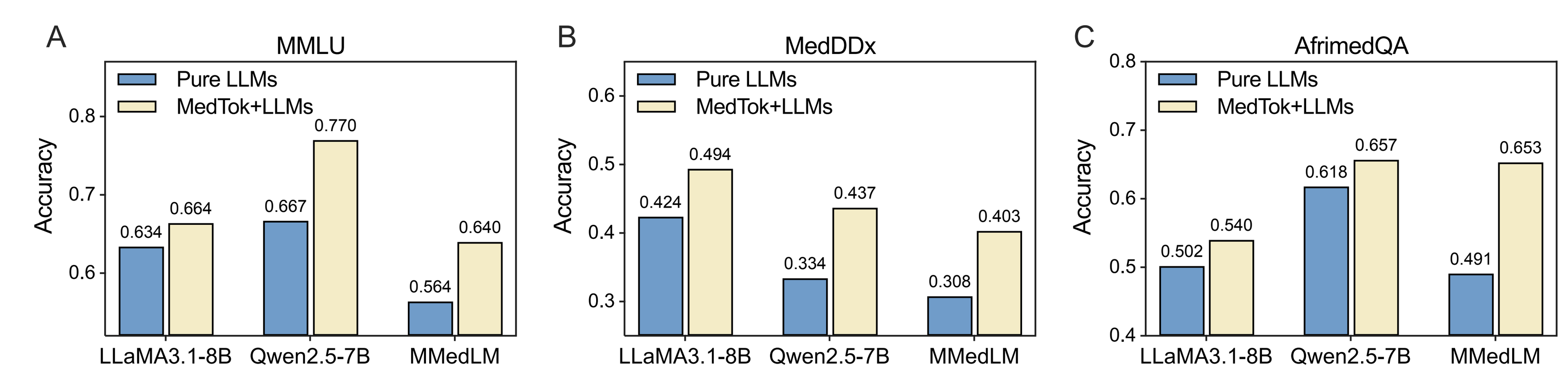


MedTok is a multimodal tokenizer that combines text descriptions of codes with relational representation of dependencies between codes. MedTok is a general-purpose tokenizer that can be used with any transformer-based model or system that requires tokenization.

MedTok for medical QA

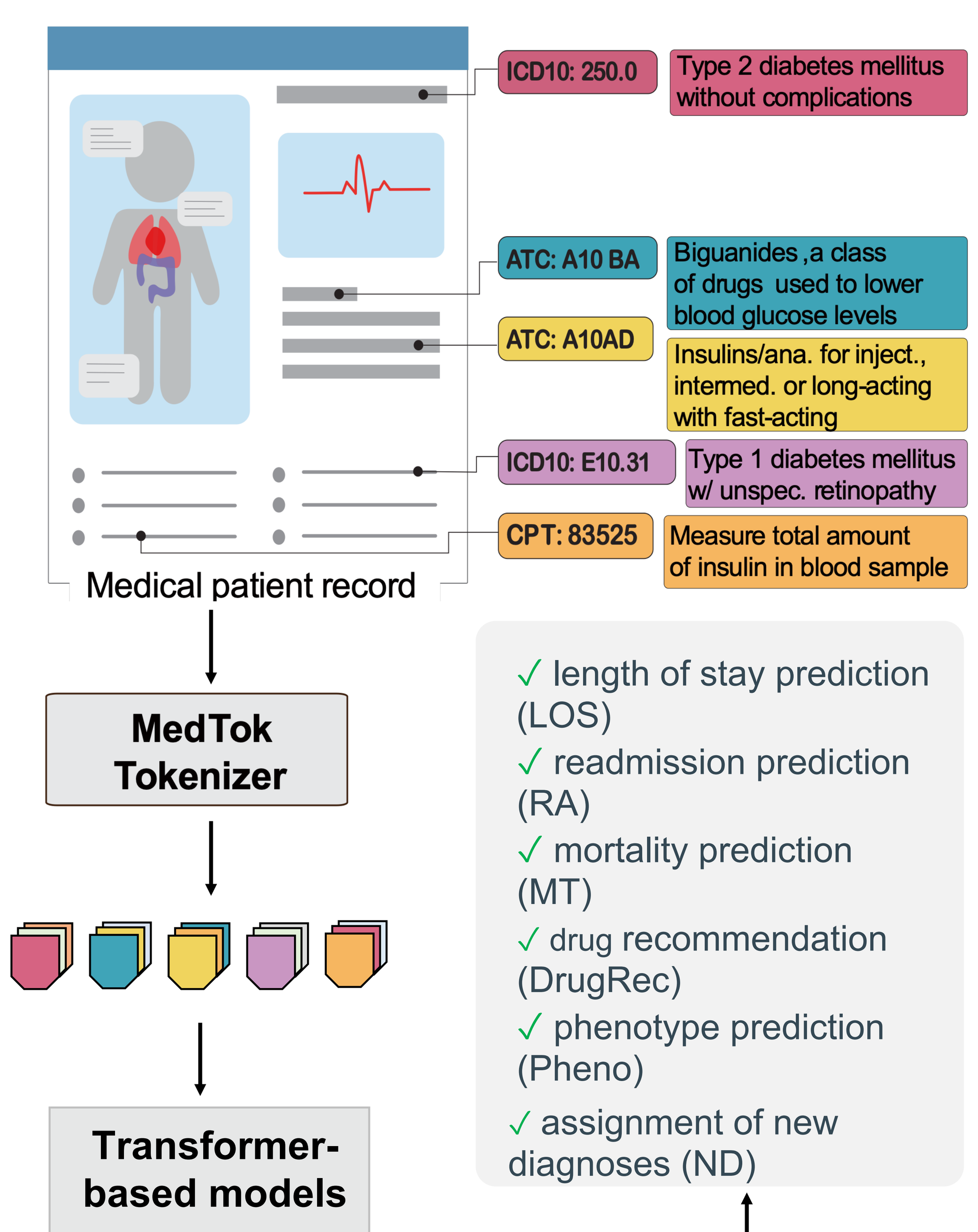


MedTok enhances few-shot learning in medical QA



- We use tokens obtained by MedTok as prefix tokens to finetune LLMs with MedMCQA dataset
- We then use other three QA datasets, including MMLU, MedDDx, and AfrimedQA, to evaluate the performances of finetuned LLMs

MedTok can be integrated into medical foundation models



- ✓ length of stay prediction (LOS)
- ✓ readmission prediction (RA)
- ✓ mortality prediction (MT)
- ✓ drug recommendation (DrugRec)
- ✓ phenotype prediction (Pheno)
- ✓ assignment of new diagnoses (ND)

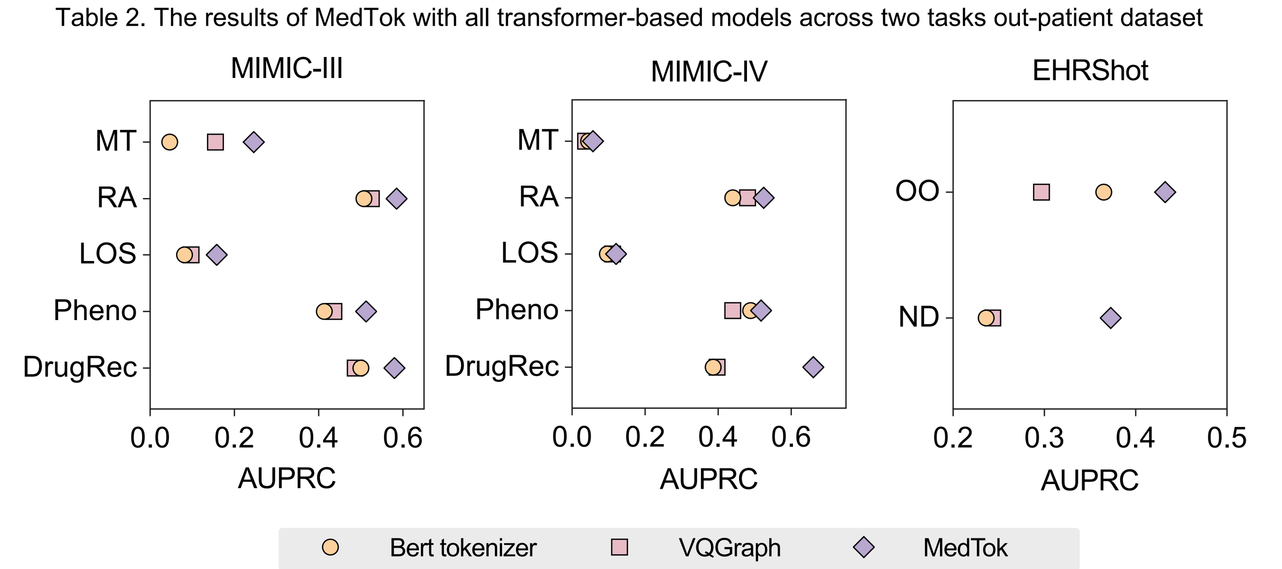
Model	Task 1: MT⁺		Task 2: RA(<15 days)⁺		Task 3: LOS*		Task 4: Pheno°		Task 5: DrugRec°	
	MIMIC-III AUPRC	MIMIC-IV AUPRC	MIMIC-III AUPRC	MIMIC-IV AUPRC	MIMIC-III AUPRC	MIMIC-IV AUPRC	MIMIC-III AUPRC	MIMIC-IV AUPRC	MIMIC-III AUPRC	MIMIC-IV AUPRC
ETHOS	0.617 (0.010)	0.282 (0.001)	0.421 (0.007)	0.648 (0.005)	N/A	N/A	N/A	N/A	0.104 (0.008)	0.131 (0.005)
+ MedTok	**0.634 (0.020)**	**0.412 (0.030)**	**0.463 (0.017)**	**0.690 (0.007)**	N/A	N/A	N/A	N/A	**0.170 (0.014)**	**0.240 (0.012)**
GT-BEHRT	0.160 (0.037)	0.028 (0.004)	0.612 (0.058)	0.586 (0.070)	0.230 (0.010)	0.103 (0.001)	0.423 (0.002)	0.493 (0.005)	0.715 (0.002)	0.736 (0.007)
+ MedTok	**0.193 (0.046)**	**0.034 (0.005)**	**0.623 (0.052)**	**0.609 (0.064)**	**0.287 (0.039)**	**0.114 (0.003)**	**0.459 (0.028)**	**0.512 (0.006)**	**0.740 (0.004)**	**0.783 (0.010)**
MulT-EHR	0.136 (0.021)	0.120 (0.003)	0.574 (0.008)	0.515 (0.007)	0.176 (0.018)	0.118 (0.032)	0.460 (0.012)	0.498 (0.001)	0.523 (0.008)	0.445 (0.027)
+ MedTok	**0.156 (0.025)**	**0.141 (0.013)**	**0.585 (0.016)**	**0.565 (0.002)**	**0.198 (0.011)**	**0.136 (0.030)**	**0.480 (0.002)**	**0.504 (0.001)**	**0.571 (0.006)**	**0.465 (0.003)**
TransformEHR	0.207 (0.012)	0.042 (0.012)	0.527 (0.030)	0.518 (0.012)	0.132 (0.021)	0.119 (0.001)	0.469 (0.022)	0.507 (0.007)	0.533 (0.030)	0.612 (0.046)
+ MedTok	**0.246 (0.044)**	**0.058 (0.007)**	**0.568 (0.036)**	**0.525 (0.017)**	**0.159 (0.031)**	**0.121 (0.002)**	**0.513 (0.024)**	**0.518 (0.012)**	**0.580 (0.035)**	**0.661 (0.092)**
BEHRT	0.163 (0.037)	0.028 (0.003)	0.529 (0.053)	0.514 (0.015)	0.232 (0.015)	0.112 (0.003)	0.587 (0.004)	0.493 (0.006)	0.539 (0.013)	0.778 (0.014)
+ MedTok	**0.220 (0.025)**	**0.032 (0.006)**	**0.574 (0.040)**	**0.515 (0.005)**	**0.251 (0.030)**	**0.137 (0.004)**	**0.603 (0.008)**	**0.504 (0.006)**	**0.558 (0.006)**	**0.792 (0.007)**
Improvement (%)	+3.32%	3.54%	3.00%	2.46%	3.13%	1.40%	2.90%	1.18%	4.10%	4.78%

+: imbalanced binary classification; ∗: multi-class classification, macro-averaged; ○: multi-label classification; N/A indicates that the model was not configured for this task.

Table 1. The results of MedTok with all transformer-based models across five tasks on two in-patient datasets

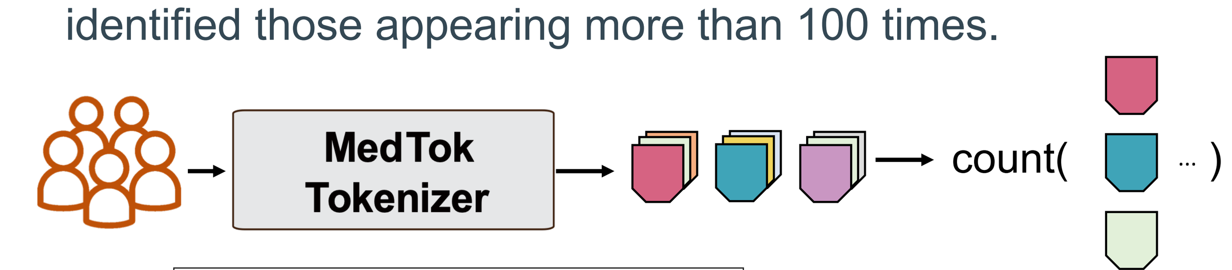
Model	Task 1: Operational Outcomes (OO)			Task 2: Assignment of New Diagnoses (ND)			
	Long LOS AUPRC	RA (<15 days) AUPRC	MT AUPRC	Hypertension AUPRC	Hyperlipidemia AUPRC	Pancreatic Cancer AUPRC	Acute MI AUPRC
ETHOS	NA	0.079 (0.017)	0.102 (0.018)	0.166 (0.020)	0.155 (0.031)	0.056 (0.006)	0.093 (0.011)
+ MedTok	NA	**0.128 (0.025)**	**0.339 (0.010)**	**0.175 (0.019)**	**0.163 (0.025)**	**0.056 (0.013)**	**0.104 (0.017)**
GT-BEHRT	0.714 (0.021)	0.115 (0.012)	0.239 (0.012)	0.303 (0.018)	0.239 (0.007)	0.044 (0.008)	0.015 (0.008)
+ MedTok	**0.739 (0.025)**	**0.154 (0.013)**	**0.444 (0.015)**	**0.360 (0.012)**	**0.441 (0.005)**	**0.074 (0.010)**	**0.031 (0.015)**
MulT-EHR	0.539 (0.025)	0.125 (0.014)	0.397 (0.016)	0.218 (0.005)	0.243 (0.005)	0.022 (0.008)	0.017 (0.003)
+ MedTok	**0.571 (0.015)**	**0.188 (0.021)**	**0.444 (0.012)**	**0.226 (0.006)**	**0.254 (0.021)**	**0.037 (0.015)**	**0.028 (0.014)**
TransformEHR	0.652 (0.023)	0.197 (0.016)	0.344 (0.030)	0.376 (0.018)	0.305 (0.021)	0.053 (0.006)	0.025 (0.006)
+ MedTok	**0.675 (0.018)**	**0.243 (0.016)**	**0.379 (0.034)**	**0.413 (0.026)**	**0.333 (0.018)**	**0.082 (0.012)**	**0.052 (0.017)**
BEHRT	0.582 (0.032)	0.332 (0.022)	0.389 (0.018)	0.233 (0.027)	0.251 (0.019)	0.036 (0.008)	0.013 (0.031)
+ MedTok	**0.723 (0.028)**	**0.397 (0.036)**	**0.431 (0.017)**	**0.287 (0.018)**	**0.302 (0.015)**	**0.057 (0.012)**	**0.036 (0.015)**
Improvement (%)	+5.52%	+5.24%	+11.32%	+3.30%	+6.00%	+1.90%	+1.76%

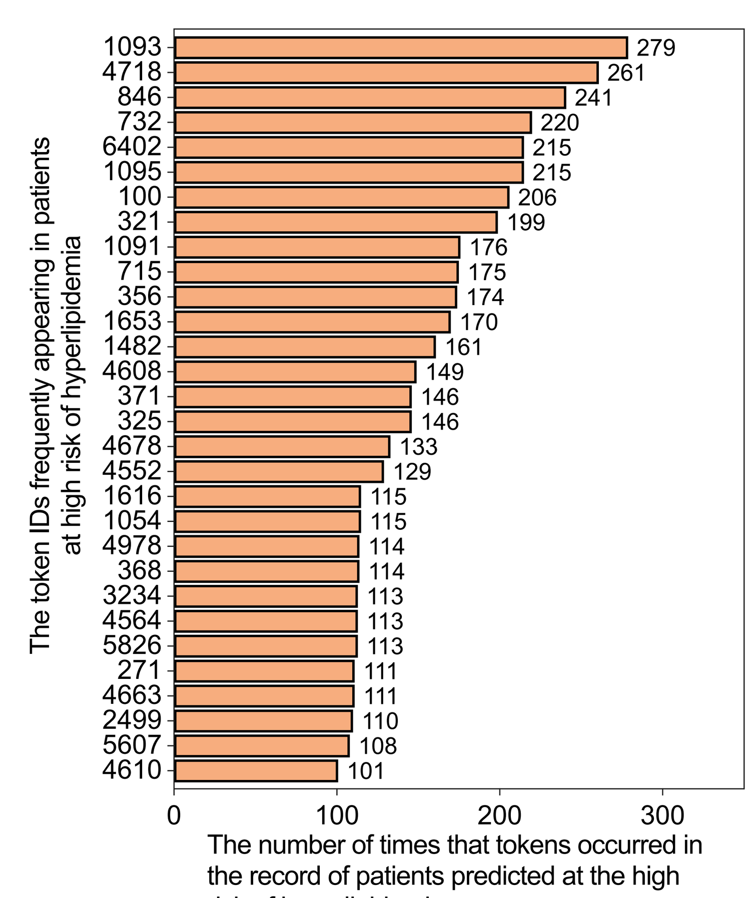
Table 2. The results of MedTok with all transformer-based models across two tasks out-patient dataset



Interpreting MedTok

- We first select patients predicted as high risk for Hyperlipidemia by MedTok with no Hyperlipidemia history.
- We then count the tokens assigned to these patients and identified those appearing more than 100 times.





We map frequently occurring tokens to medical codes. The most frequent codes are:

- ✓ Rosuvastatin 5mg Oral Tablet
- ✓ Burn of skin
- ✓ Type 2 diabetes mellitus without complication (disorder)
- ✓ Hyperlipidemia

Try out MedTok

```
from transformers import AutoTokenizer

tokenizer = AutoTokenizer.from_pretrained("mims-harvard/MedTok", trust_remote_code=True)
tokens = tokenizer("E11.9")
embed = tokenizer.embed("E11.9")
```

HuggingFace Paper Code