DKEC: Domain Knowledge Enhanced Multi-Label Classification for Electronic Health Records

Xueren Ge, Abhishek Satpathy, Ronald Dean Williams, John A. Stankovic, Homa Alemzadeh

University of Virginia, Charlottesville, VA 22903 USA

{zar8jw, cqa3ym, rdw, jas9f, ha4d}@virginia.edu

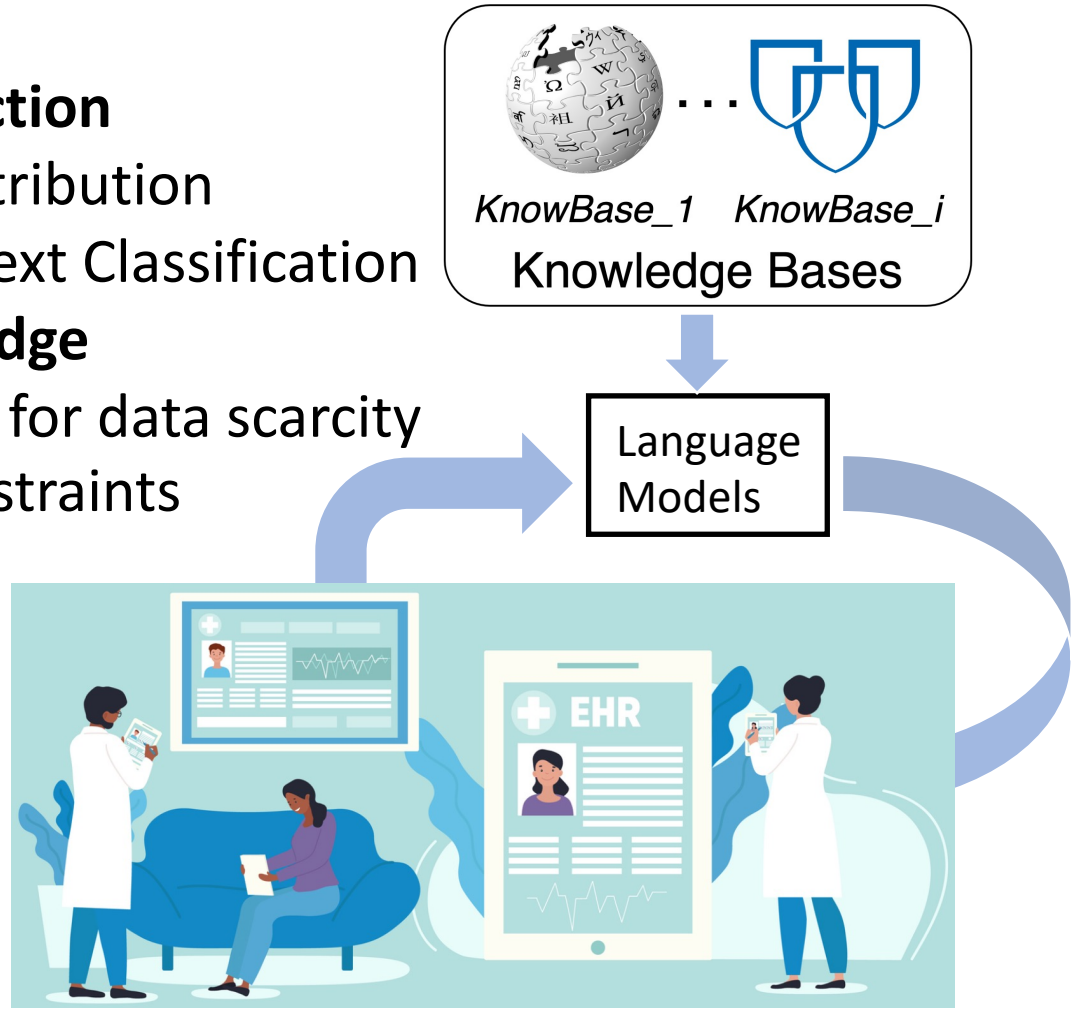
UNIVERSITY of VIRGINIA ENGINEERING

LINK LAB

UVA DSA

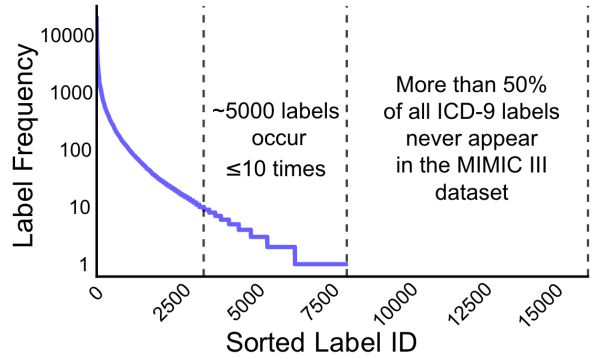
Electronic Health Records

- **Diagnosis prediction**
  - Long-tail Distribution
  - Multi-label text Classification
- **Domain Knowledge**
  - Compensate for data scarcity
  - Relation constraints



Diagnosis Prediction
(Multi-label text classification)

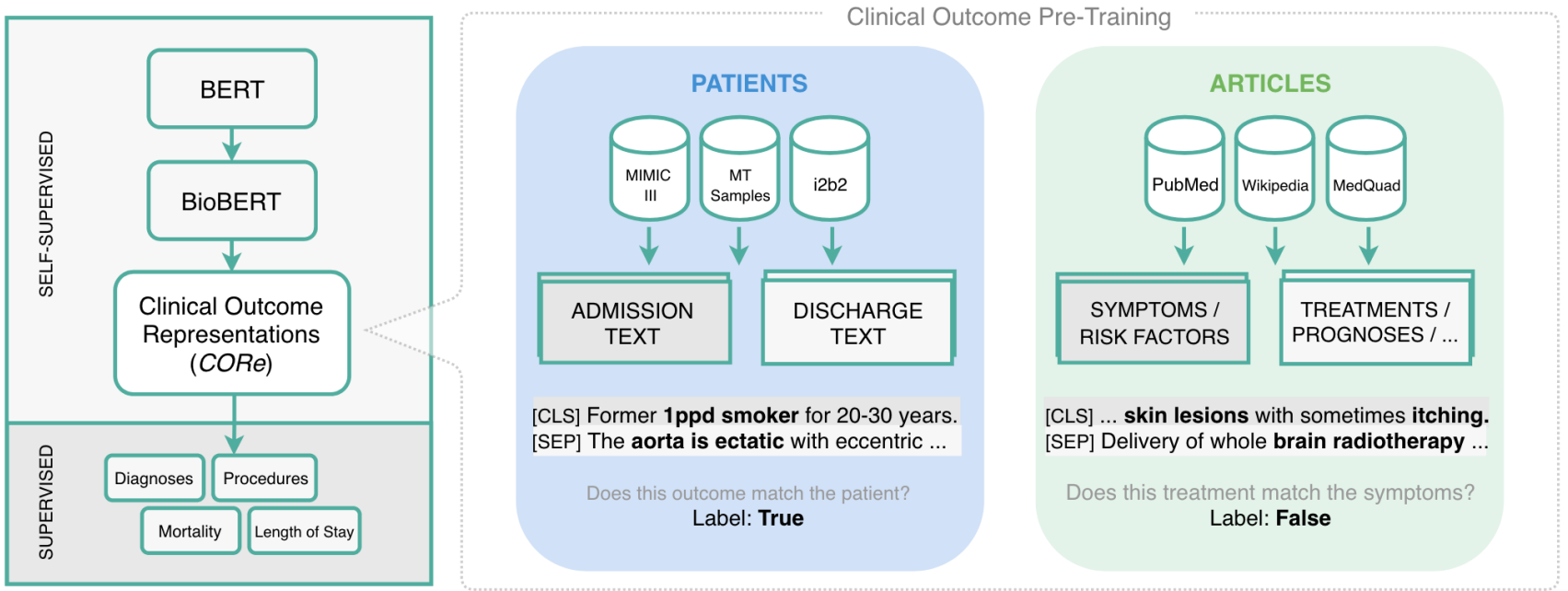
- ICD-9 diagnose code
- EMS Protocols



Rios, Anthony, and Ramakanth Kavuluru. "Few-shot and zero-shot multi-label learning for structured label spaces." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. Vol. 2018. NIH Public Access, 2018.

Related Work

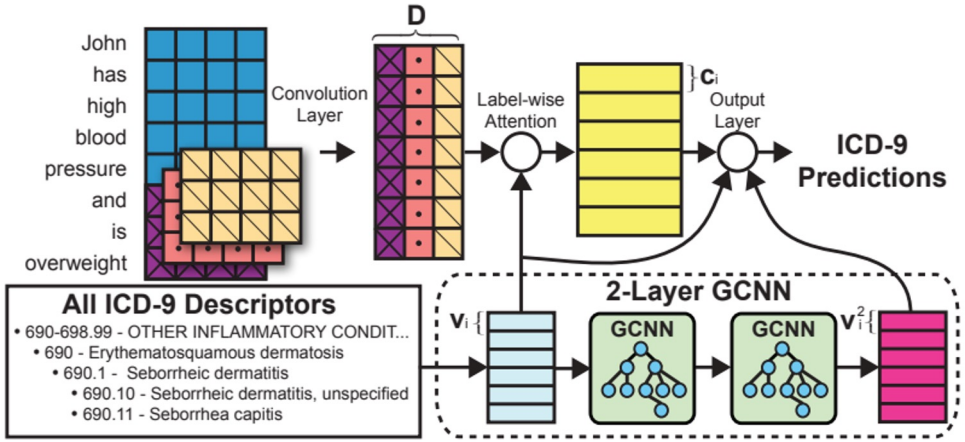
- Pretrained language models for diagnosis prediction
  - BioBERT[1](110M), GatorTron[2](325M), BioMedLM[3](2.7B)

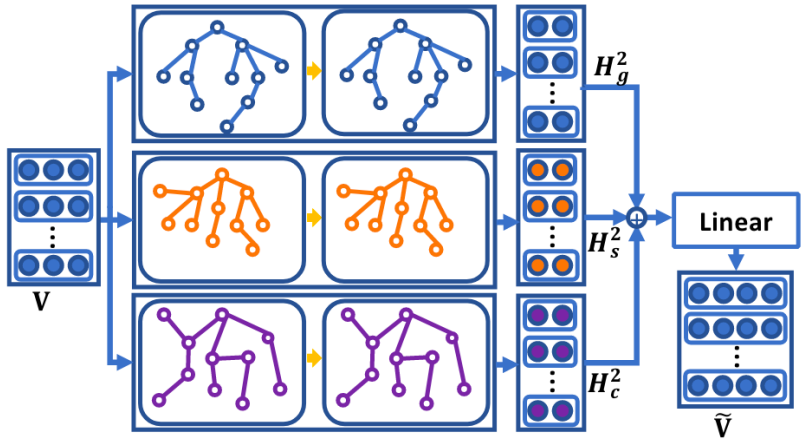


[1] Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36.4 (2020): 1234-1240.
[2] Yang, Xi, et al. "GatorTron: A Large Clinical Language Model to Unlock Patient Information from Unstructured Electronic Health Records." arXiv preprint arXiv:2203.03540 (2022).
[3] Bolton Elliot, Hall David, Yasunaga Michihiro, Lee Tony, Manning Chris, and Liang Percy. 2022. Biomedlm.

Related Work

- Knowledge enhanced approaches
  - **Semantic meaning of labels** was integrated into text embeddings in CAML[1]
  - **Hierarchical label structures** was encoded into a graph model and further concatenated into text features in ZAGCNN[2]
  - **Label semantics similarity**, **label co-occurrence** along with **label hierarchy** was utilized to construct multiple graphs in KAMG[3]





[1] Mullenbach, James, et al. "Explainable prediction of medical codes from clinical text." arXiv preprint arXiv:1802.05695 (2018).
[2] Rios, Anthony, and Ramakanth Kavuluru. "Few-shot and zero-shot multi-label learning for structured label spaces." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. Vol. 2018. NIH Public Access, 2018.
[3] Lu, Jueqing, et al. "Multi-label few/zero-shot learning with knowledge aggregated from multiple label graphs." arXiv preprint arXiv:2010.07459 (2020).

Related Work

- Common Limitations of Existing Works

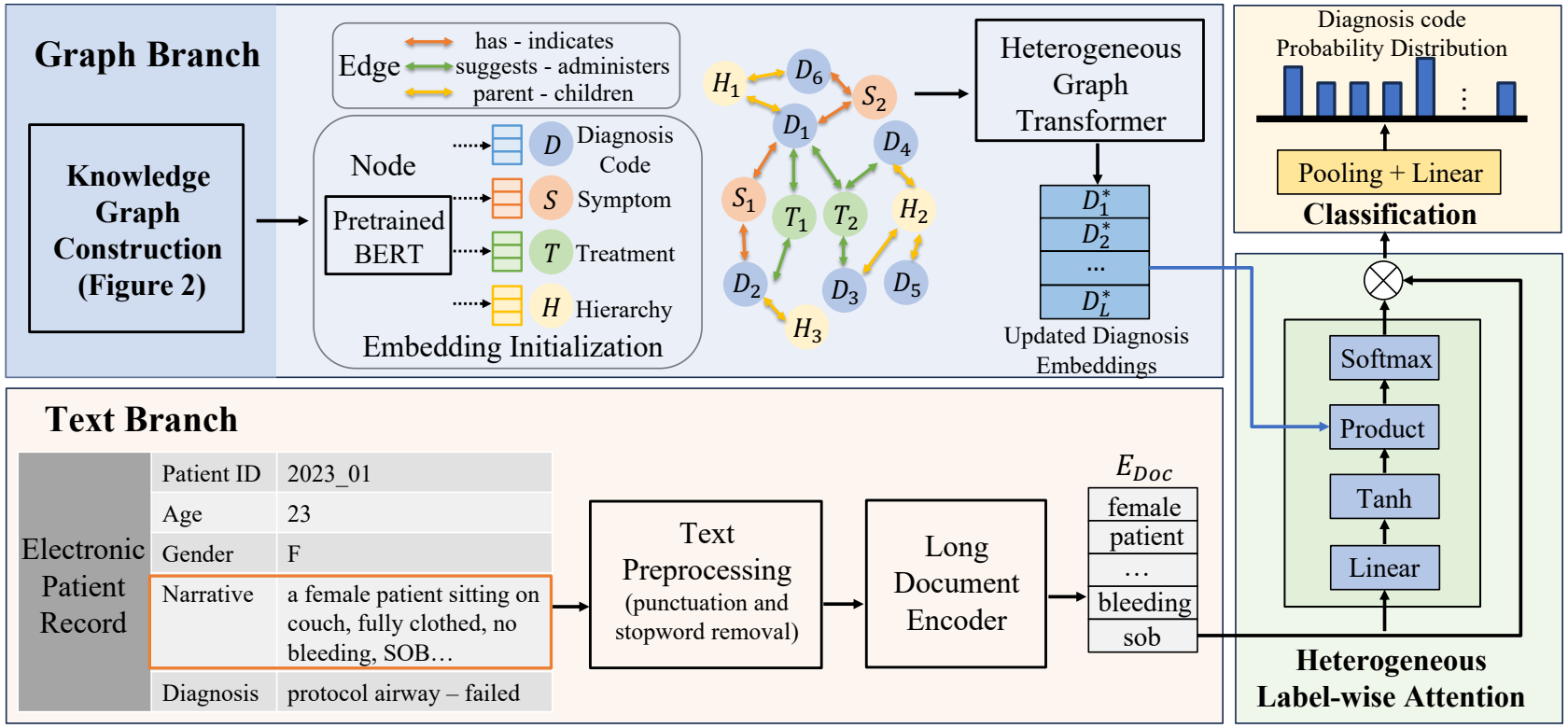
Models	Encoder	Attention Mechanism	Knowledge Integration	Knowledge Source	Datasets
(van Aken et al., 2021b)	BERT	Self-Attention	Pre-training	Wikipedia, PubMed	MIMIC-III
(Yang et al., 2022b)	MegatronBERT	Self-Attention	Pre-training	Wikipedia, PubMed	MIMIC-III
(Bolton et al., 2024)	GPT2	Self-Attention	Pre-training	PubMed	MedMCQA
(Mullenbach et al., 2018)	CNN	Label-wise Attention	ICD-9 hierarchy graph	ICD-9 description	MIMIC-III
(Rios and Kavuluru, 2018)	CNN	Label-wise Attention			MIMIC-III
(Li and Yu, 2020)	Multi-filter residual CNN	Label-wise Attention			MIMIC-III
(Zhou et al., 2021)	Multi-filter CNN	Shared Interactive Attention			MIMIC-III
DKEC (Ours)	Multi-filter CNN, Transformers	Label-wise Attention	Heterogeneous graph	Wikipedia, MayoClinic, ODEMSA	MIMIC-III & EMS

Table 1: Summary of previous works on diagnosis prediction.

- Contribution
  - Knowledge graph construction by **GPT-4 with chain-of-thought prompting**.
  - Knowledge graph incorporation with language models by **heterogeneous label-wise attention** to improve multi-label classification
  - DKEC outperform SOTAs on two real-world datasets and enables smaller language models achieve comparable performance to large language models.

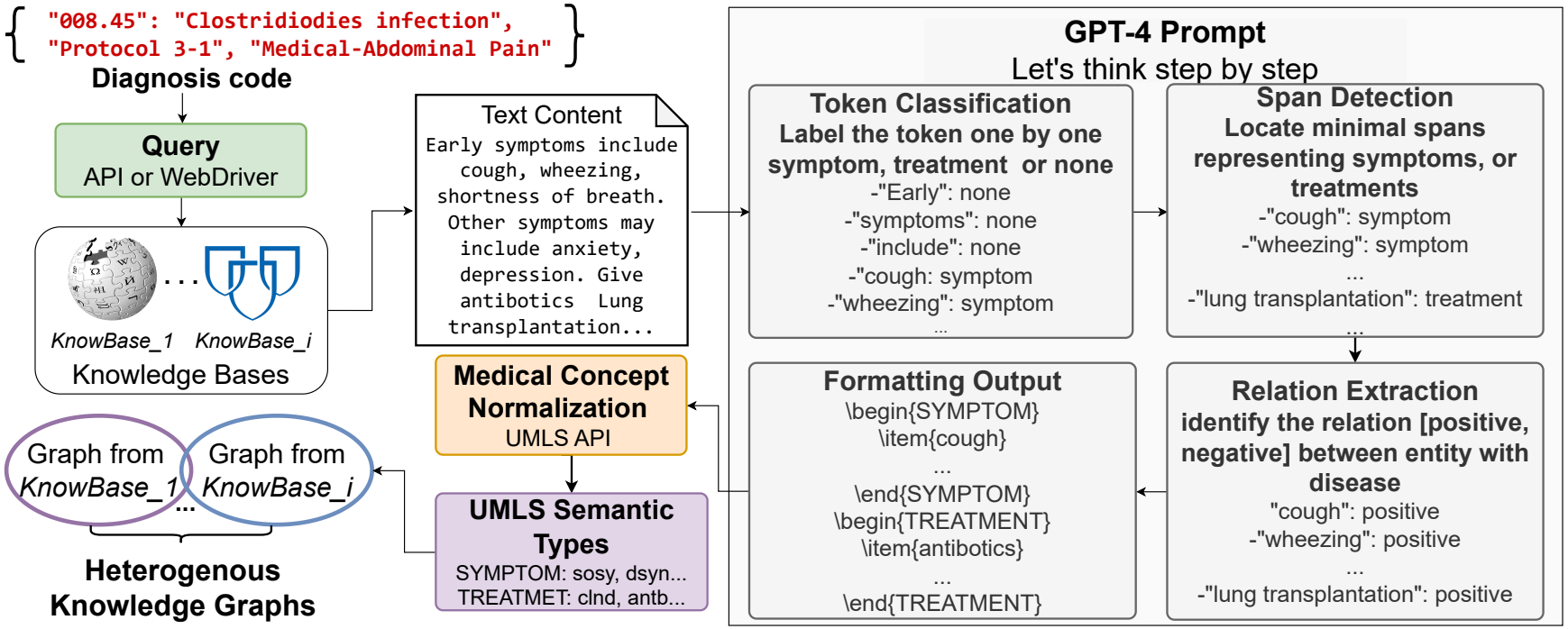
DKEC pipeline

- Knowledge Graph Construction
- Graph Branch
- Text Branch
- Heterogeneous Label-wise Attention
- Classification



Knowledge Graph Construction

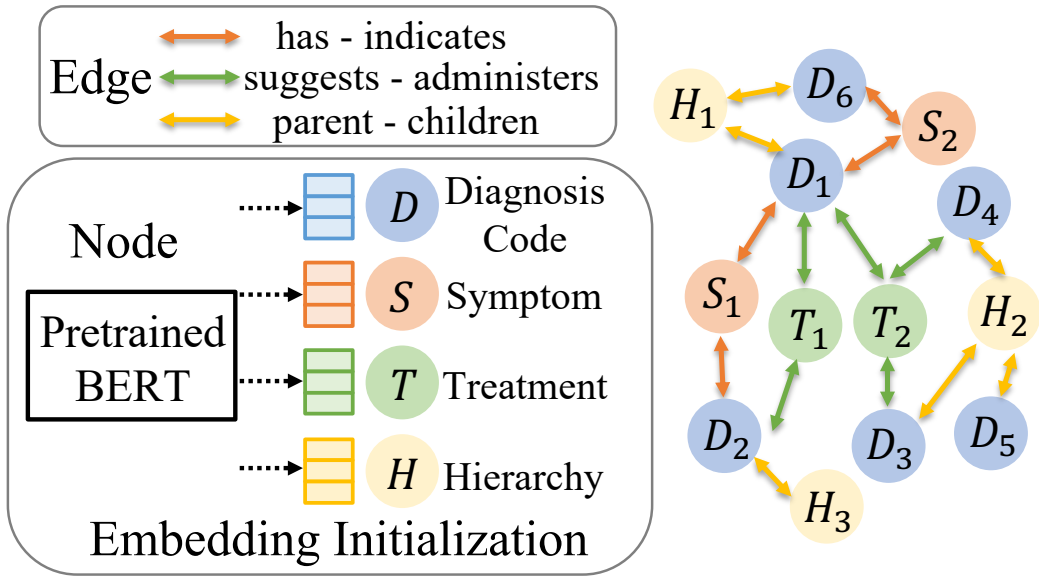
- Information query
- GPT-4 CoT prompt
  - Token classification
  - Span Detection
  - Relation Extraction
- UMLS normalization
- Union of Multiple Knowledge Bases



Heterogeneous Graph

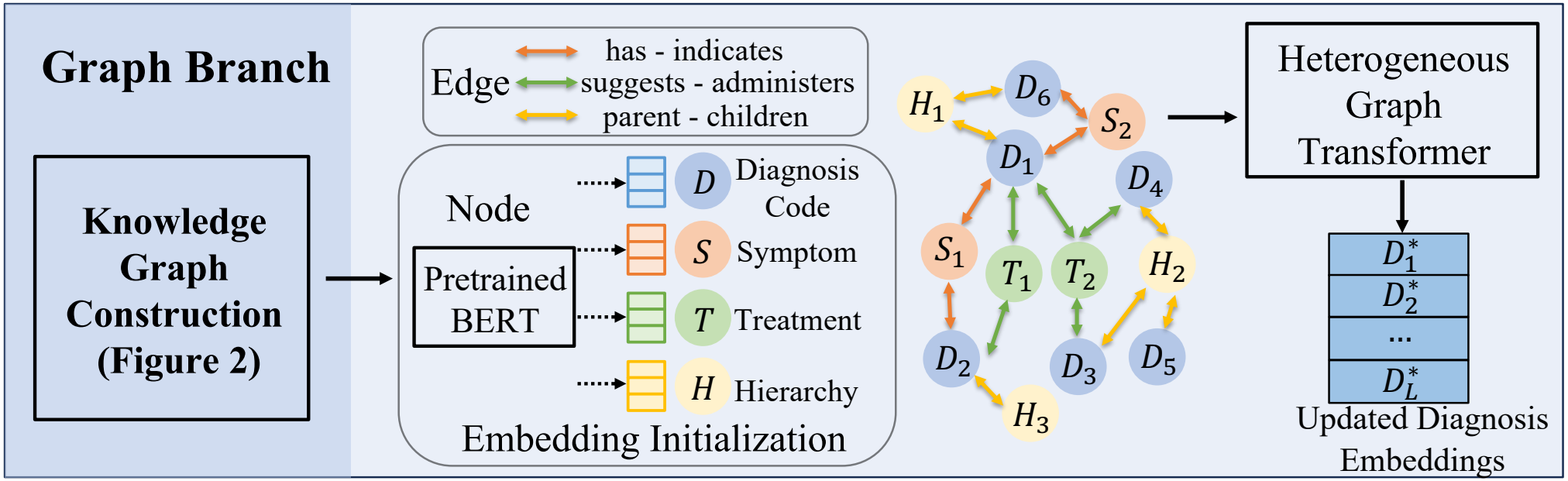
- Nodes
  - Diagnose code $D = \{D_k\}_{k=1}^{L}$
  - Signs and Symptoms $S = \{S_k\}_{k=1}^{|S|}$
  - Procedures $T = \{T_k\}_{k=1}^{|T|}$
  - Hierarchy $H = \{H_k\}_{k=1}^{|H|}$
- Edges
  - Diagnosis Code – Signs: $\overleftrightarrow{E_{DS}}$
  - Diagnosis Code –Treatment: $\overleftrightarrow{E_{DS}}$
  - Diagnosis Code – Hierarchy: $\overleftrightarrow{E_{DH}}$

- Embedding Initialization
  - Pre-trained BERT
  - Diagnose code: Overview
  - Other medical entities: Names



Graph Branch

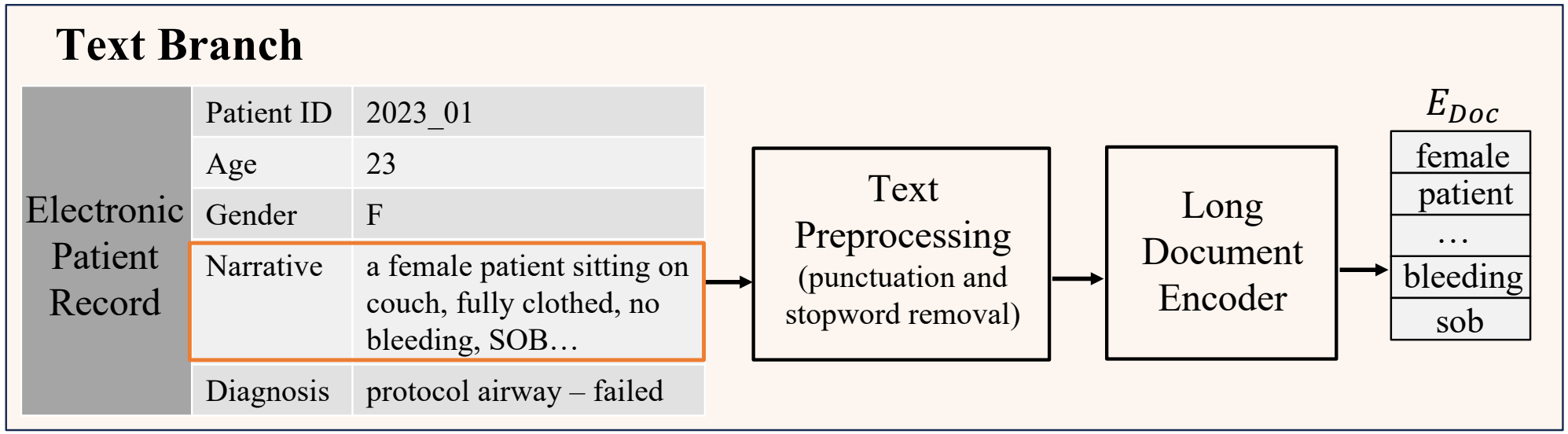
- Input: Knowledge Graph with initial node embedding
- Graph Model: Heterogeneous Graph Transformer[1]
- Output: Updated Diagnosis code embedding D_k^*



[1] Hu, Ziniu, et al. "Heterogeneous graph transformer." Proceedings of the web conference 2020. 2020.

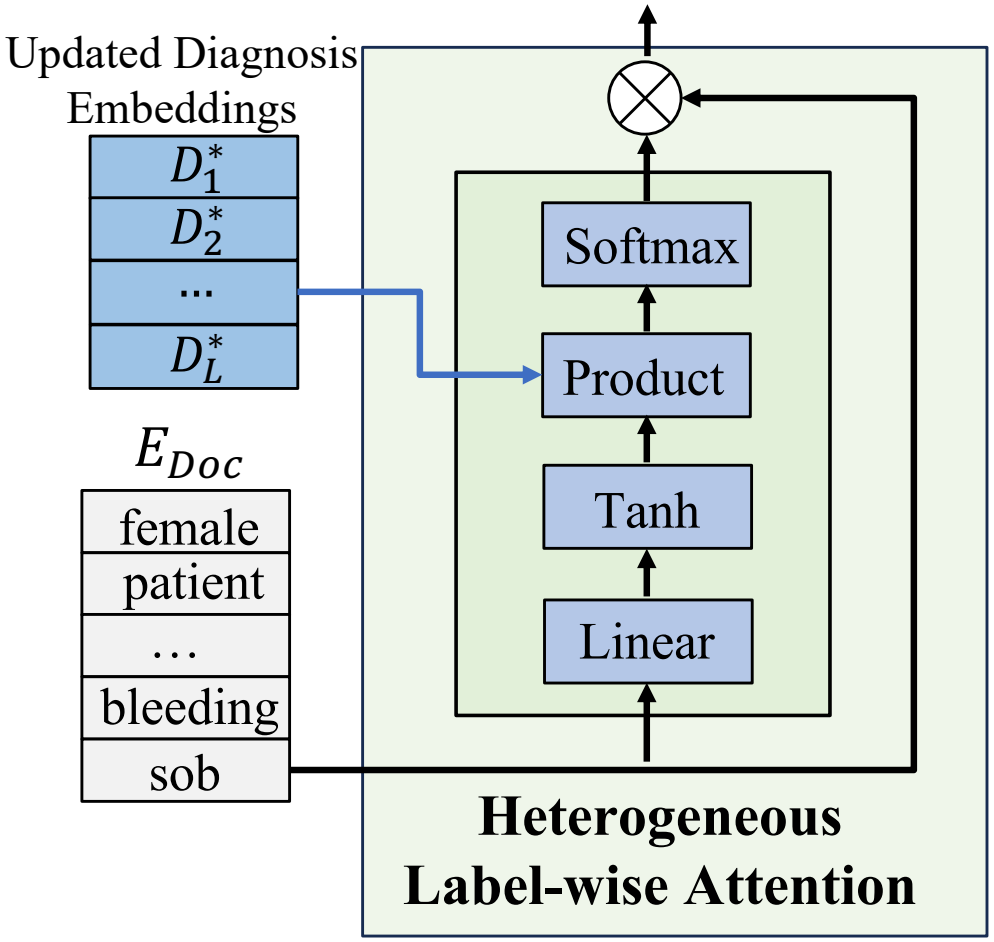
Text Branch

- Long Document Encoder
  - Multi-filter CNNs
  - Pre-trained Transformers
  - Document representation E_{Doc}



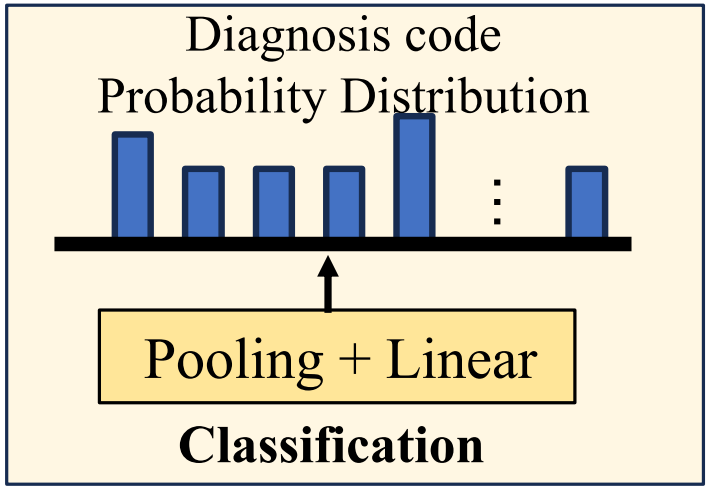
Heterogeneous Label-wise Attention

- Given document representation E_{Doc} and kth label representation D_k^*, we derived kth label-wise attention vectors $a_{Doc,k}$, which **having kth label assign different weights for each token in the document representation**.
  - $a_{Doc,k} = softmax(tanh(W_0 E_{Doc} + b_0) D_k^*)$
- Combine all label-wise attention vectors
  - $A_{Doc} = [a_{Doc,1}\ a_{Doc,2}\ \dots\ a_{Doc,k}\ \dots\ a_{Doc,L}\ ]^T$
- Apply label-wise attention vectors to text representation, the label-attentive text representation which measures how informative medical document $\boldsymbol{Doc}$ is for all L labels
  - $E_{Doc}^{attn} = A_{Doc} E_{Doc}$



Classification

- Pooling and Linear Layer
  - $\hat{y}_{Doc} = \text{Linear}(\text{Pooling}(E_{Doc}^{attn}))$
- Binary cross-entropy Loss
  - $\mathcal{L} = \sum_{Doc=1}^{|Doc|} \sum_{l=1}^{L} y_{Doc,l} \log(\hat{y}_{Doc,l}) + (1 - y_{Doc,l}) \log(1 - \hat{y}_{Doc,l})$



Experiments

- We apply DKEC to different baseline language models and compare it with state-of-the-art label-wise attention networks and large language models, we aim to answer three research questions:
  - ***RQ1:*** **Can DKEC improve MLTC performance for class-imbalanced datasets?**
  - ***RQ2:*** **How does DKEC perform when applied to language models with varying sizes?**
  - ***RQ3:*** **How does DKEC perform with scaling label sizes?**

Knowledge Bases (KBs) and Datasets

- Emergency Medical Service (EMS):
  - A collection of 4,417 pre-hospital ePCR annotated with EMS protocols
  - KBs: ODEMSA documents
- MIMIC-III:
  - A real-world EHR dataset annotated with ICD-9 diagnosis codes
  - KBs: Wikipedia & Mayo Clinic web
- Label Frequency
  - Head(**H**): labels with more than **100** samples
  - Middle(**M**): labels with **10** to **100** samples
  - Tail(**T**): labels with less than **10** samples

Dataset	N_{train}	N_{val}	N_{test}	N_l H	N_l M	N_l T
EMS	2787	314	1316	10	21	12
MIMIC-III	47413	1627	3363	494	1038	2205

Table 3: Dataset statistics, N_{train}: number of training instances, N_{val}: number of validation instances, N_{test}: number of test instances, N_l: number of labels in total.

[1] Kim, Sion, et al. "Information Extraction from Patient Care Reports for Intelligent Emergency Medical Services." 2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). IEEE, 2021.
[2] Cao, Pengfei, et al. "HyperCore: Hyperbolic and co-graph representation for automatic ICD coding." Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

Knowledge Graph Quality Evaluation

- Manual annotation of Symptoms and treatments
  - 50 ICD-9 diagnosis codes (evenly sampled from head, middle, tail labels)
  - 43 EMS protocols

- One-shot CoT GPT-4 outperforms other baselines in medical entity extraction consistently by a considerable margin

wo/w NORM	Wikipedia (50 ICD-9 codes)		Mayo Clinic (50 ICD-9 codes)		ODEMSA (43 EMS protocols)	
	Symptom	Treatment	Symptom	Treatment	Symptom	Treatment
MetaMap	47.62 / 51.53	34.66 / 41.95	44.83 / 49.12	41.82 / 46.44	41.34 / 43.61	39.20 / 41.95
cTAKES	48.74 / 52.58	36.01 / 43.35	42.60 / 46.67	39.67 / 45.35	38.02 / 42.47	48.96 / 52.31
ScispaCy	52.79 / 55.57	41.73 / 49.71	46.54 / 50.43	45.94 / 50.89	44.39 / 47.69	35.88 / 38.82
zero-shot GPT-4	51.99 / 58.77	17.93 / 32.13	52.98 / 63.37	26.16 / 36.48	76.07 / 79.72	10.17 / 23.50
one-shot CoT GPT-4	**84.63 / 86.57**	**85.70 / 89.12**	**82.03 / 86.72**	**90.43 / 93.90**	**86.96 / 91.01**	**86.48 / 88.92**

Table 2: Comparison with baselines on three knowledge bases. *wo/w NORM* means the micro F1-scores are measured before/after medical entity normalization. The best results are highlighted in **bold**.

RQ1: Can DKEC improve MLTC performance for class-imbalanced datasets?

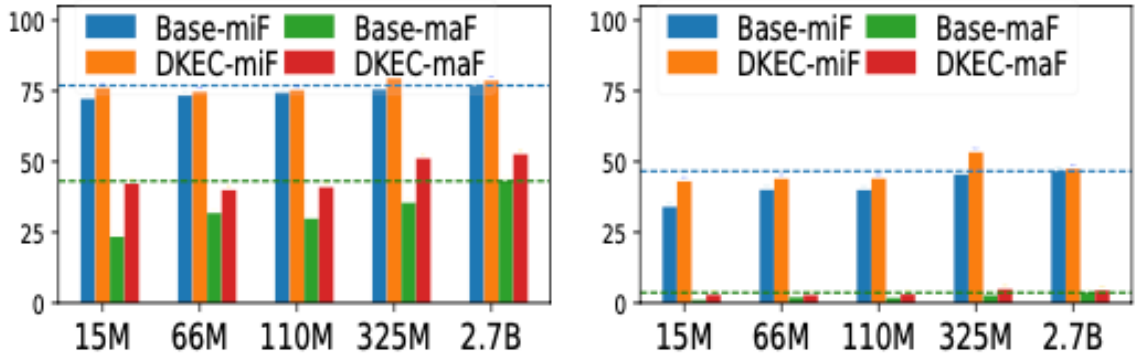
		Head Labels		Middle Labels		Tail Labels		Overall			
		P@1	R@1	P@1	R@1	P@1	R@1	miF	maF	P@1	R@1
EMS	CAML	$78.6_{\pm 1.3}$	$77.7_{\pm 1.3}$	$33.0_{\pm 0.5}$	$32.6_{\pm 0.6}$	$22.7_{\pm 4.5}$	$22.7_{\pm 4.5}$	$63.7_{\pm 1.2}$	$22.4_{\pm 1.3}$	$65.0_{\pm 1.6}$	$63.5_{\pm 1.5}$
	ZAGCNN	$83.0_{\pm 1.0}$	$82.0_{\pm 1.0}$	$47.0_{\pm 1.0}$	$46.2_{\pm 0.7}$	$37.9_{\pm 7.7}$	$37.9_{\pm 7.7}$	$64.8_{\pm 1.1}$	$28.3_{\pm 2.0}$	$69.6_{\pm 0.7}$	$68.1_{\pm 0.6}$
	MultiResCNN	$84.3_{\pm 0.2}$	$83.2_{\pm 0.2}$	$35.6_{\pm 1.8}$	$35.0_{\pm 2.0}$	$25.0_{\pm 2.3}$	$25.0_{\pm 2.3}$	$65.8_{\pm 0.2}$	$26.1_{\pm 0.5}$	$67.9_{\pm 0.3}$	$66.3_{\pm 0.3}$
	ISD	$81.7_{\pm 0.9}$	$80.8_{\pm 0.9}$	$44.2_{\pm 0.4}$	$43.2_{\pm 0.5}$	$29.5_{\pm 2.3}$	$29.5_{\pm 2.3}$	$67.1_{\pm 1.2}$	$26.1_{\pm 0.1}$	$68.0_{\pm 1.3}$	$66.5_{\pm 1.2}$
	GatorTron	$\underline{89.4}_{\pm 0.5}$	$\underline{88.4}_{\pm 0.5}$	$66.0_{\pm 0.4}$	$64.7_{\pm 0.7}$	$\underline{57.1}_{\pm 2.2}$	$\underline{57.1}_{\pm 2.2}$	$75.5_{\pm 0.6}$	$35.4_{\pm 1.9}$	$77.3_{\pm 0.6}$	$75.4_{\pm 0.6}$
	BioMedLM	$89.3_{\pm 0.3}$	$88.2_{\pm 0.3}$	$\underline{71.3}_{\pm 0.7}$	$\underline{70.1}_{\pm 0.6}$	$47.6_{\pm 4.3}$	$47.6_{\pm 4.3}$	$\underline{76.9}_{\pm 0.7}$	$\underline{43.1}_{\pm 1.7}$	$\underline{78.4}_{\pm 0.6}$	$\underline{76.6}_{\pm 0.6}$
	DKEC-M-CNN	$85.2_{\pm 0.7}$	$83.0_{\pm 0.7}$	$53.2_{\pm 1.3}$	$52.7_{\pm 1.1}$	$45.1_{\pm 2.1}$	$45.1_{\pm 2.1}$	$68.6_{\pm 0.4}$	$32.4_{\pm 0.6}$	$72.4_{\pm 0.4}$	$71.7_{\pm 0.6}$
	DKEC-GatorTron	$\mathbf{91.8}_{\pm 0.1}$	$\mathbf{90.7}_{\pm 0.1}$	$\mathbf{72.4}_{\pm 0.4}$	$\mathbf{71.3}_{\pm 0.4}$	$\mathbf{67.6}_{\pm 2.3}$	$\mathbf{67.6}_{\pm 2.3}$	$\mathbf{79.5}_{\pm 0.5}$	$\mathbf{51.1}_{\pm 1.5}$	$\mathbf{82.2}_{\pm 0.5}$	$\mathbf{80.3}_{\pm 0.6}$
		P@8	R@8	P@8	R@8	P@8	R@8	miF	maF	P@8	R@8
MIMIC-III	CAML	$54.8_{\pm 0.5}$	$57.5_{\pm 0.6}$	$5.5_{\pm 0.4}$	$28.4_{\pm 2.3}$	$0.7_{\pm 0.1}$	$4.8_{\pm 0.5}$	$51.5_{\pm 0.7}$	$4.3_{\pm 0.5}$	$54.4_{\pm 0.5}$	$50.3_{\pm 0.5}$
	ZAGCNN	$55.3_{\pm 0.2}$	$58.0_{\pm 0.2}$	$6.6_{\pm 0.1}$	$34.4_{\pm 0.7}$	$1.8_{\pm 0.1}$	$11.7_{\pm 0.8}$	$52.1_{\pm 0.4}$	$4.0_{\pm 0.3}$	$55.2_{\pm 0.2}$	$51.2_{\pm 0.3}$
	MultiResCNN	$\underline{56.5}_{\pm 0.3}$	$\underline{59.4}_{\pm 0.2}$	$\underline{8.2}_{\pm 0.5}$	$\underline{42.3}_{\pm 2.8}$	$1.2_{\pm 0.1}$	$7.5_{\pm 0.9}$	$\mathbf{55.6}_{\pm 0.3}$	$\mathbf{6.0}_{\pm 0.6}$	$\underline{56.6}_{\pm 0.2}$	$\underline{52.7}_{\pm 0.2}$
	ISD	$51.8_{\pm 0.5}$	$53.8_{\pm 0.5}$	$6.1_{\pm 0.2}$	$31.7_{\pm 1.2}$	$1.9_{\pm 0.2}$	$12.6_{\pm 0.9}$	$46.8_{\pm 1.3}$	$2.8_{\pm 0.2}$	$51.6_{\pm 0.5}$	$47.5_{\pm 0.5}$
	GatorTron	$50.4_{\pm 0.2}$	$53.4_{\pm 0.2}$	$6.5_{\pm 0.2}$	$33.8_{\pm 1.1}$	$2.0_{\pm 0.3}$	$12.7_{\pm 1.4}$	$45.4_{\pm 0.4}$	$2.7_{\pm 0.3}$	$50.3_{\pm 0.2}$	$47.1_{\pm 0.2}$
	BioMedLM	$50.5_{\pm 0.1}$	$53.4_{\pm 0.1}$	$6.1_{\pm 0.1}$	$31.3_{\pm 1.2}$	$\underline{2.0}_{\pm 0.1}$	$\underline{13.2}_{\pm 1.1}$	$46.6_{\pm 0.3}$	$3.7_{\pm 0.5}$	$50.2_{\pm 0.1}$	$47.2_{\pm 0.2}$
	DKEC-M-CNN	$\mathbf{58.6}_{\pm 0.2}$	$\mathbf{61.5}_{\pm 0.2}$	$\mathbf{9.6}_{\pm 0.1}$	$\mathbf{49.2}_{\pm 0.8}$	$2.9_{\pm 0.1}$	$\mathbf{19.2}_{\pm 0.9}$	$\underline{55.0}_{\pm 0.3}$	$4.9_{\pm 0.2}$	$\mathbf{58.9}_{\pm 0.2}$	$\mathbf{54.8}_{\pm 0.2}$
	DKEC-GatorTron	$56.8_{\pm 0.4}$	$59.8_{\pm 0.2}$	$8.5_{\pm 0.1}$	$44.7_{\pm 0.7}$	$\mathbf{3.1}_{\pm 0.2}$	$19.1_{\pm 1.1}$	$53.0_{\pm 0.4}$	$\underline{5.7}_{\pm 0.3}$	$56.9_{\pm 0.4}$	$53.2_{\pm 0.3}$

Table 4: Comparison with SOTA on EMS and MIMIC-III (**RQ1**). The best and runner-up results are in **bold** and underlined.

RQ2: How does DKEC perform when applied to language models with varying sizes?

- Performance of DKEC-based models increase less as model size grows

- DKEC enables smaller language models to achieve comparable performance to LLMs



(a) EMS (b) MIMIC-III

Figure 3: DKEC with different sizes of pre-trained transformers (**RQ2**).

RQ3: How does DKEC perform with scaling label sizes?

- With the increase in the number of labels, the MLTC performance generally drops, but DKEC helps maintain performance, particularly when external knowledge for all the labels.

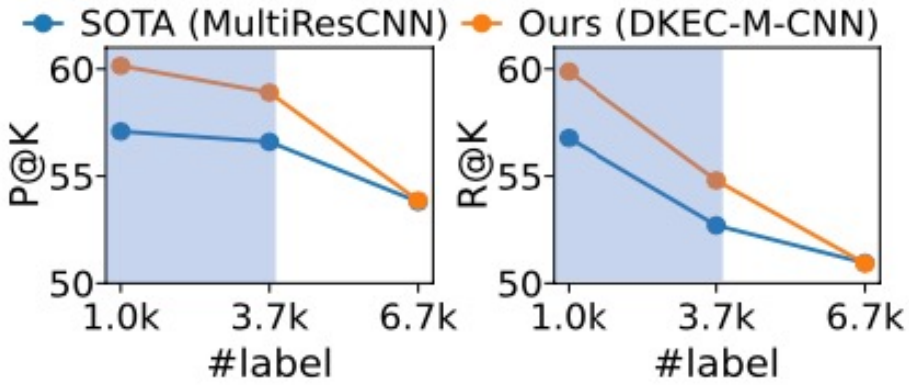


Figure 4: Performance on subsets of MIMIC-III dataset with varying label sizes (**RQ3**). Subsets with 1.0k and 3.7k labels have full knowledge, and 6.7k has partial knowledge.

Ablation Study

- Effectiveness of DKEC
  - DKEC outperforms SOTAs that uses only label hierarchy
- Effectiveness of External Knowledge
  - Incorporating label-specific external knowledge is the main driver of performance improvements.

Encoder	Label-wise Attention	miF	maF	P@8	R@8
1-CNN	Label hierarchy*	52.1	4.0	55.2	51.2
1-CNN	DKEC	54.8	4.2	57.5	53.3
GatorTron	Label hierarchy	46.6	3.2	50.7	47.5
GatorTron	DKEC	53.0	**5.7**	56.9	53.2
M-CNN	DKEC	55.0	4.9	**58.9**	**54.8**
M-CNN	DKEC w/o hierarhcy	**55.2**	4.9	58.6	54.5

Table 5: Ablation study using MIMIC-III dataset. "1/M-CNN" are the single/multi-filter CNN. 1-CNN with Label hierarchy* represents the SOTA ZAGCNN.

Limitations

- Only 3,737 MIMIC-III diagnosis codes are found having domain knowledge from Wikipedia and Mayo Clinic, larger KBs may be needed for completeness.
- The accuracy of the full knowledge graph would require a considerable amount of human effort.
- DKEC and other baselines studied in this paper can only used as a reference and not as a final decision for treatment of the patients in the real world.

Abhishek Satpathy
cqa3ym@virginia.edu

Ronald Dean Williams
rdw@virginia.edu

John A. Stankovic
stankovic@cs.virginia.edu

Homa Alemzadeh
ha4d@virginia.edu

Thank You!

zar8jw@virginia.edu

National Institute of Standards and Technology
U.S. Department of Commerce

This work was supported by the award 70NANB21H029 from the U.S. Department of Commerce, National Institute of Standards and Technology (NIST)

Github: https://github.com/UVA-DSA/DKEC