MKeCL: Medical Knowledge-Enhanced Contrastive Learning for Few-shot Disease Diagnosis

Yutian Zhao*, Huimin Wang*, Xian Wu†, Yefeng Zheng
Jarvis Research Center, Tencent YouTu Lab
Shenzhen, China
{yutianzhao, hmmmwang, kevinxwu, yefengzheng}@tencent.com

Background and Motivation

1) long tail problem:
there are excessive EMRs for common diseases and insufficient EMRs for rare diseases, thus training over an imbalanced data set could result in a biased model that ignores rare diseases in diagnosis.

2) easily misdiagnosed diseases:
some diseases can be easily distinguished while others sharing analogous conditions are much more difficult.

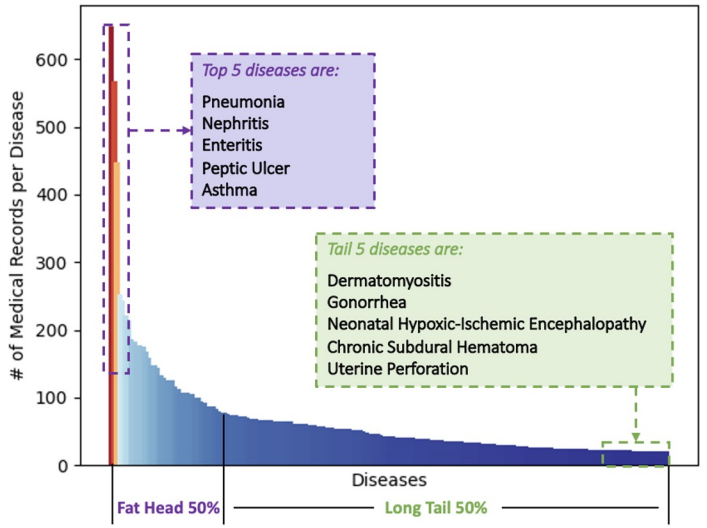

Figure 1: Distribution of 12,776 electronic medical records collected from five hospitals. Only half of the diseases have more than 40 associated records and less than 10% of diseases have more than 100 records.

Methodology: Medical Knowledge-Enhanced Contrastive Representation Learning (MKeCL)

1) Medical knowledges are converted into question-answer pairs

2) The Transformer encoder is used to generate embeddings for each question-answer pair.

3) Prediction: embeddings of all disease candidates are compared with that of EMR, and the one with the smallest cosine distance is the predicted disease.

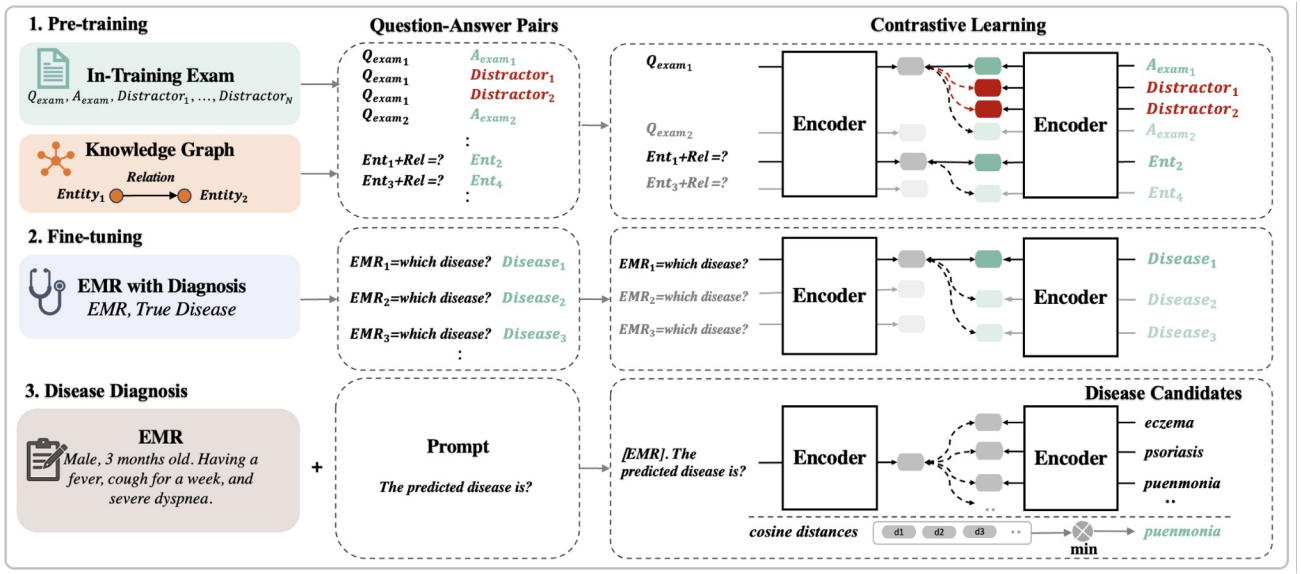

Figure 2. The Workflow of MKeCL. — between embeddings represents positive instances, - - means in-batch negatives, and - - denote hard negatives.

Dataset

Pretraining

- Medical Knowledge Graph
  - 2,585 disease-related triples from a pre-built medical KG
- Medical Licensing Exam
  - 41,626 multiple-choice questions

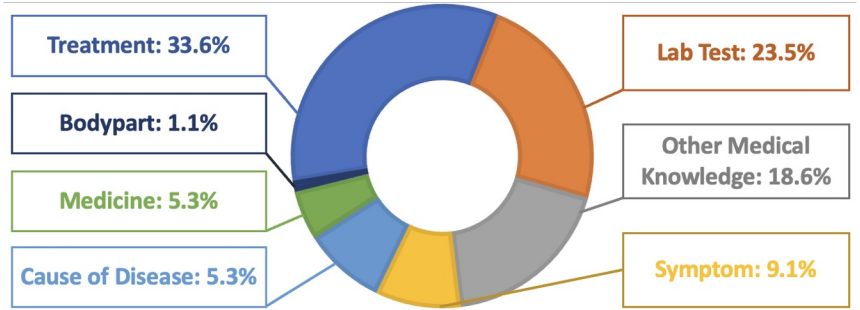

Figure 3. Distribution of different types of medical knowledge used in pretraining MKeCL.

Finetuning

- Electronic Medical Records
  - 12,776 electronic medical records that cover 193 diseases

Question: Male, middle-aged. Papules and rashes appear all over the body. What disease does he have?
Options: A Psoriasis. B Eczema. C Dermatitis. D Skin cancer.
Correct answer is A

Question: Male, middle-aged. Repeating diarrhea and stomachache. What is the cause of disease?
Options: A Tuberculosis. B Parasitosis. C Immune response. D Viral infection.
Correct answer is C

Question: Female, elderly. Experienced high fever and sudden abdominal pain and chills. What lab test should be done?
Options: A MRCP. B Abdominal ultrasound. C Abdominal CT scan. D ERCP.
Correct answer is B

Figure 4. Examples of multiple–choice questions in Medical Licensing Exam. Each question contains only one correct answer and the rest are distractors.

Metrics

Micro F1

$$\text{micro} - \text{F}\,1 = 2\frac{\text{Recall}_\text{m} \times \text{P}\,\text{recision}_\text{m}}{\text{Recall}_\text{m} + \text{P}\,\text{recision}_\text{m}}$$

Macro F1

$$Macro\text{-}F_1 = \frac{1}{|T|}\sum_{t \in T}\frac{2P_tR_t}{P_t + R_t}$$

Alignment

$$\ell_{uniform} \triangleq log \underset{(x,y) \overset{i.i.d}{\sim} p_{data}}{\mathbb{E}} e^{-2||f(x)-f(y)||^2}$$

Uniformity

$$\ell_{align} \triangleq \underset{(x,x^+) \sim p_{pos}}{\mathbb{E}} ||f(x) - f(x^+)||^2$$

Main Results

Model	0%		1%		3%		5%		10%		100%	
	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma	Mi	Ma
ALBERT (Lan et al., 2019)	-	-	30.9	31.8	36.1	43.2	48.3	44.7	61.6	57.6	87.7	85.9
BERT (Devlin et al., 2018)	-	-	42.6	44.1	43.8	50.7	55.4	52.1	68.6	63.2	90.3	**89.6**
MedBERT (Ting et al., 2020)	-	-	42.2	43.5	44.2	50.9	54.9	52.4	67.0	61.9	90.2	89.1
GP (Yang et al., 2022a)	-	-	43.5	52.8	49.2	59.0	48.7	55.8	54.8	58.5	82.7	85.6
KEPT (Yang et al., 2022b)	-	-	47.4	45.1	51.9	54.6	60.2	53.8	68.7	63.2	89.8	87.9
ChatGPT	40.9	41.2	41.1	40.5	43.0	44.6	45.2	43.5	45.3	44.8	-	-
GPT-4	41.7	34.4	40.6	32.4	44.1	36.2	46.3	37.7	46.2	37.3	-	-
MKeCL$_{mlm\ pretrain}$	-	-	50.6	43.56	51.2	49.4	58.1	50.2	68.4	60.5	89.7	86.0
MKeCL$_{w/o\ pretrain}$	-	-	45.2	39.9	49.2	49.2	55.2	45.3	66.5	59.1	88.3	84.1
MKeCL$_{w/o\ exam}$	23.7	16.6	48.3	44.3	52.4	51.6	56.8	47.1	67.2	59.2	88.9	85.5
MKeCL$_{w/o\ kg}$	49.5	46.1	59.3	54.4	60.4	58.3	65.4	57.5	71.5	64.0	90.2	87.8
MKeCL	**50.5**	**46.1**	**60.7**	**55.7**	**63.7**	**60.5**	**68.1**	**61.6**	**73.0**	**67.0**	**90.5**	87.4

Table 1: The micro F1 (Mi) and macro F1 (Ma) on EMR dataset. We sample 0%, 1%, 3%, 5%, and 10% of the dataset to train models respectively and evaluate their performance at zero-shot and few-shot settings. For ChatGPT and GPT-4, we adopt in-context learning to simulate the few-shot setting. This involves demonstrating 1, 2, 3, and 5 examples in the prompt to mimic the few-shot setup.

Visualization of Medical Record Representations

Visualization of medical record representations generated by ALBERT, BERT, MedBERT, KEPT, and MKeCL using t-SNE. Models are trained on 1% of the training dataset and medical records from 10 diseases are sampled randomly for visualization.

- 10 diseases: Nephritis, Anemia, Peptic, Ulcer, Enteritis, Tuberculuois, SLE, Bone Fracture, Pancreatitis, Asthma

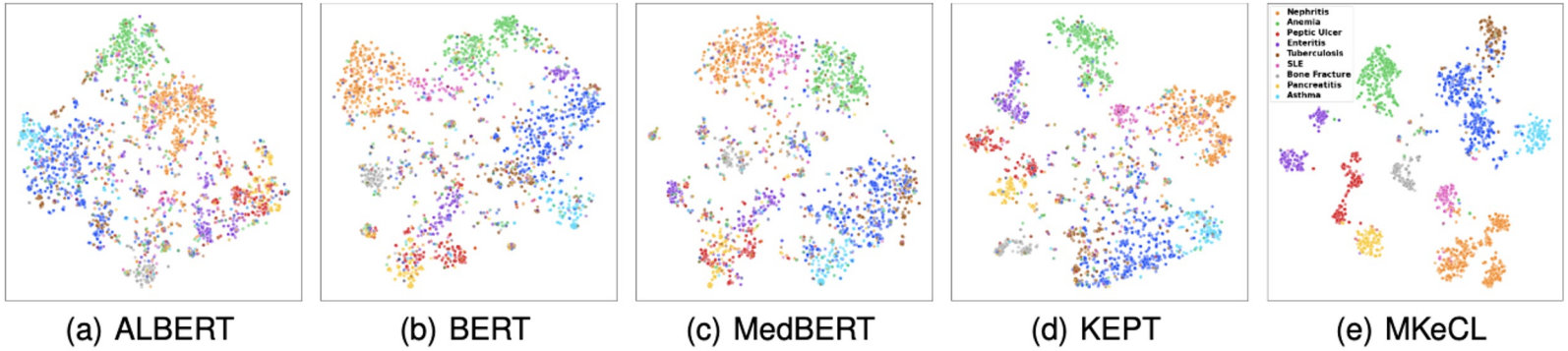

Figure 5. Visualization of medical record representations.

Visualization of Medical Record Representations

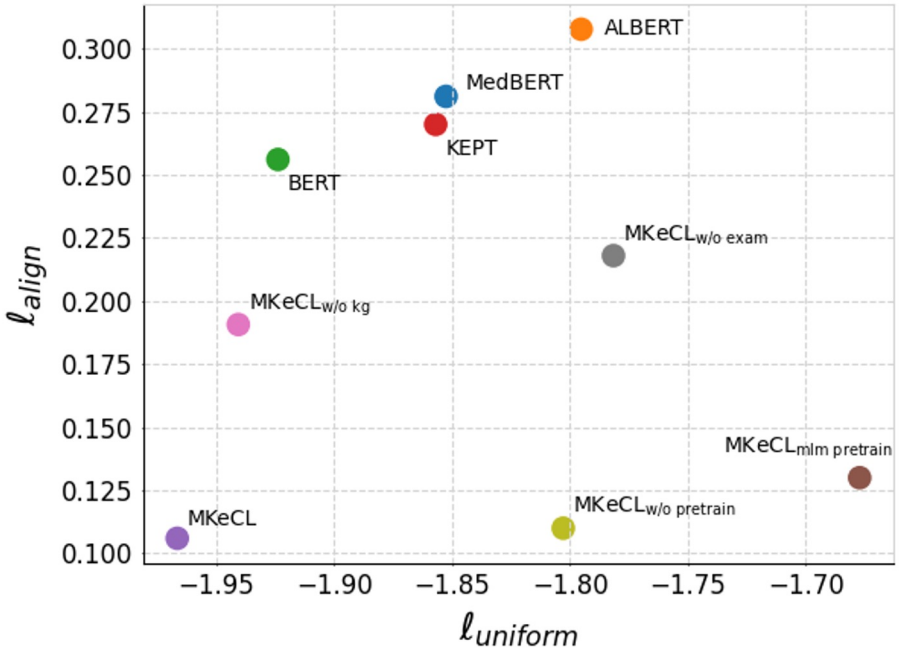

Figure 6. Plot of alignment and uniform for all baseline models and all variants of MKeCL

Comparison of Easily Misdiagnosed Diseases

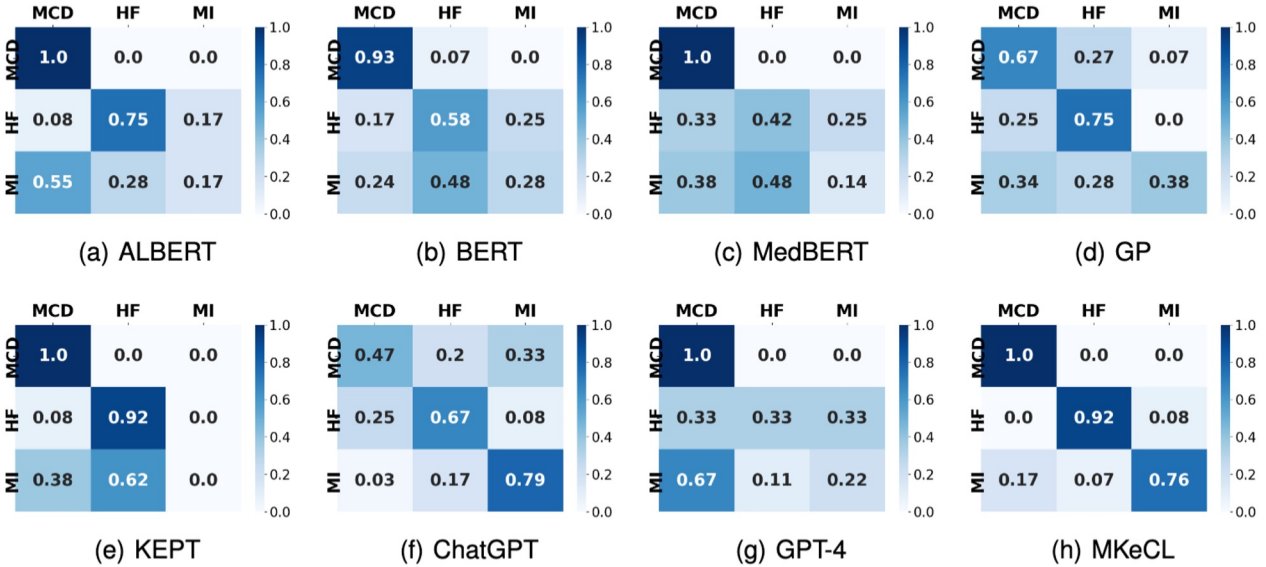

Figure 7: Confusion matrix on easily misdiagnosed diseases. We use a heatmap to visualize each model's accuracy in classifying between Myocarditis (MCD), Heart Failure (HF), and Myocardial Infarction (MI). The x-axis represents the correct disease and the y-axis represents the predicted disease.

Case Study on Easily Misdiagnosed Diseases

Case 1: Male, elderly, with a 3-year history of exertional angina. Over the past 2 weeks, the frequency of angina episodes has increased, and blood pressure has risen to 166/94 mmHg. The patient also experiences paroxysmal nocturnal dyspnea and is currently unable to lie flat.

Case 2: Female, middle-aged, with persistent chest pain for 6 hours. Physical examination: BP 110/70 mmHg. No crackles or wheezes were detected in both lungs. Heart rate is 125 beats/minute with a regular rhythm. No murmurs are heard in any cardiac valve areas. The electrocardiogram reveals partial ST-T elevation. Laboratory tests indicate elevated blood troponin levels.

EMR	GT	BERT	ALBERT	MedBERT	GP	KEPT	ChatGPT	GPT-4	MKeCL
Case 1	HF	HF	HF	MCD	MCD	HF	HF	HF	HF
Case 2	MCD	HF	MCD	MCD	HF	MI	MI	MI	MCD

Table 2. Disease prediction results of different algorithms on two real cases. MCD, HF, and MI are Myocarditis, Heart Failure, and Myocardial Infarction, respectively

Thanks for listening!