EHRKit: A Python Natural Language Processing Toolkit for Electronic Health Record Texts

Rui Yang*, Yujie Qiao*, Qingcheng Zeng*, Keen You, Xiangru Tang, Lucas Huang, Chia-Chun Hsieh, Benjamin Rosand, Jeremy Goldwasser, Dragomir Radev, Irene Li#

Introduction

🔍 **Research Motivation:**

- As Electronic Health Records (EHRs) become increasingly prevalent, massive unstructured texts are generated within the healthcare system. The secondary usage of these unstructured texts holds great importance, but the main obstacle is the processing and understanding of them.
- There are existing libraries and toolkits are designed for biomedical needs, including Stanza, SciFive, UmlsBERT, MIMIC-Extract and so on. However, there is a need for new toolkit which can cover a wider range of clinical NLP tasks. Moreover, the exploration of generative tasks for unstructured texts within EHRs remains limited.

🚀 **Research Contributions:**

- **EHRKit:** We propose EHRKit, a Python NLP toolkit for EHR unstructured texts. This toolkit contains two main components: general API functions and MIMIC-specific functions. It is user-friendly, with easy installation and quick start tutorials.
- To address the gap in generative tasks for clinical unstructured texts, we also develop machine translation, summarization, understandable text translation, and chatbot functions in clinical scenarios based on existing pre-trained models.

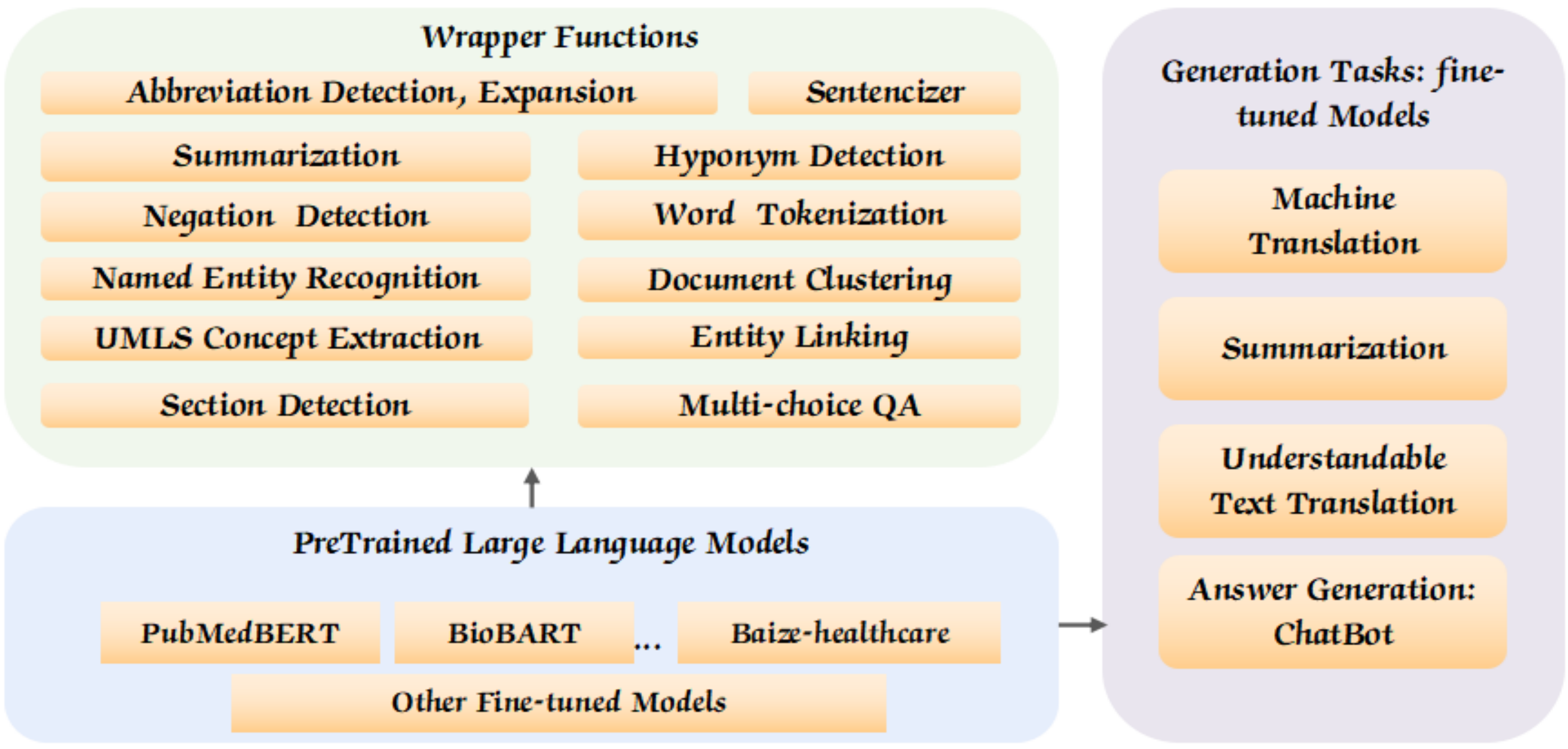
Evaluation

📊 **Benchmarks:**

- **Machine Translation:** UFAL Medical Corpus.
- **Summarization:** Three public medical datasets: PubMed, MIMIC-CXR, MEDQA.
- **Understandable Text Translation:** Three public medical datasets: MedLane, eLife, PLOS.
- **Question-Answering:** In Multiple-choice QA part, two public medical datasets: HEADQA and MedMCQA. In Answer Generation part, MedQUAD dataset is used for finetuning and test questions of the TREC-2017 LiveQA medical task are used for evaluation.

🔬 **Evaluation:**

- We use already existing models such as BART, Pegasus, BIOBART, Baize, PubMedBERT etc. for finetuning and testing.



Wrapper Functions

Abbreviation Detection, Expansion

Sentencizer

Summarization

Hyponym Detection

Negation Detection

Word Tokenization

Named Entity Recognition

Document Clustering

UMLS Concept Extraction

Entity Linking

Section Detection

Multi-choice QA

PreTrained Large Language Models

PubMedBERT

BioBART

...

Baize-healthcare

Other Fine-tuned Models

Generation Tasks: fine-tuned Models

Machine Translation

Summarization

Understandable Text Translation

Answer Generation: ChatBot

Fig 1. EHRKit Architecture.

	MIMIC	Neu	MT	Summ	UTT	Chat
MIMIC-Extract	✓					
ScispaCy		✓				
medspaCy		✓				
Stanza Biomed		✓				
SciFive		✓		✓		
EHRKit	✓	✓	✓	✓	✓	✓

Tab 1. A comparison with other similar python toolkits.
MIMIC: MIMIC Related. Neu: Neural Methods. MT: Machine Translation.
Summ: Summarization. UTT: Understandable Text Translation. Chat: Chatbot.

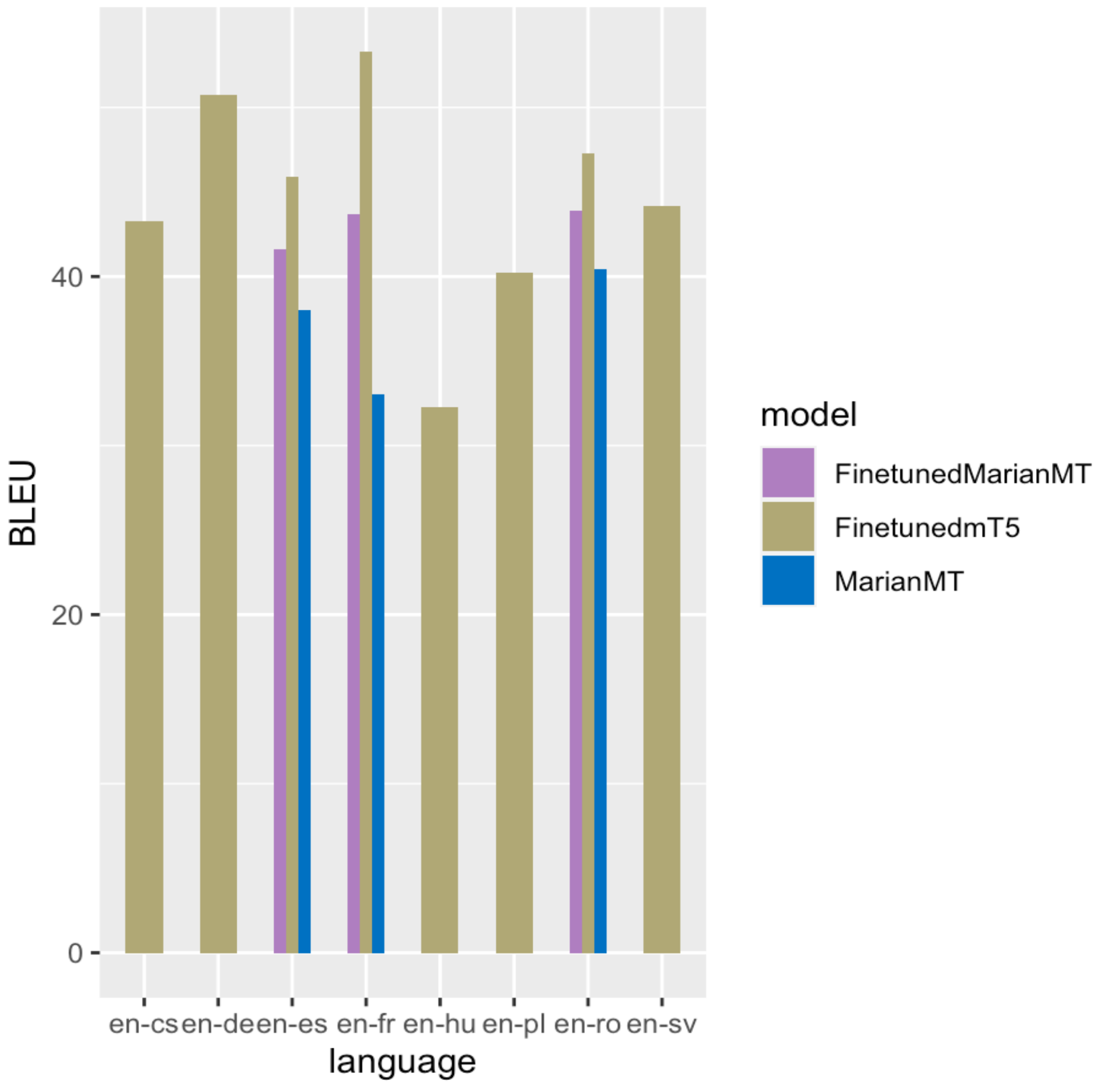


Fig 2. Evaluation of Machine Translation Tasks.

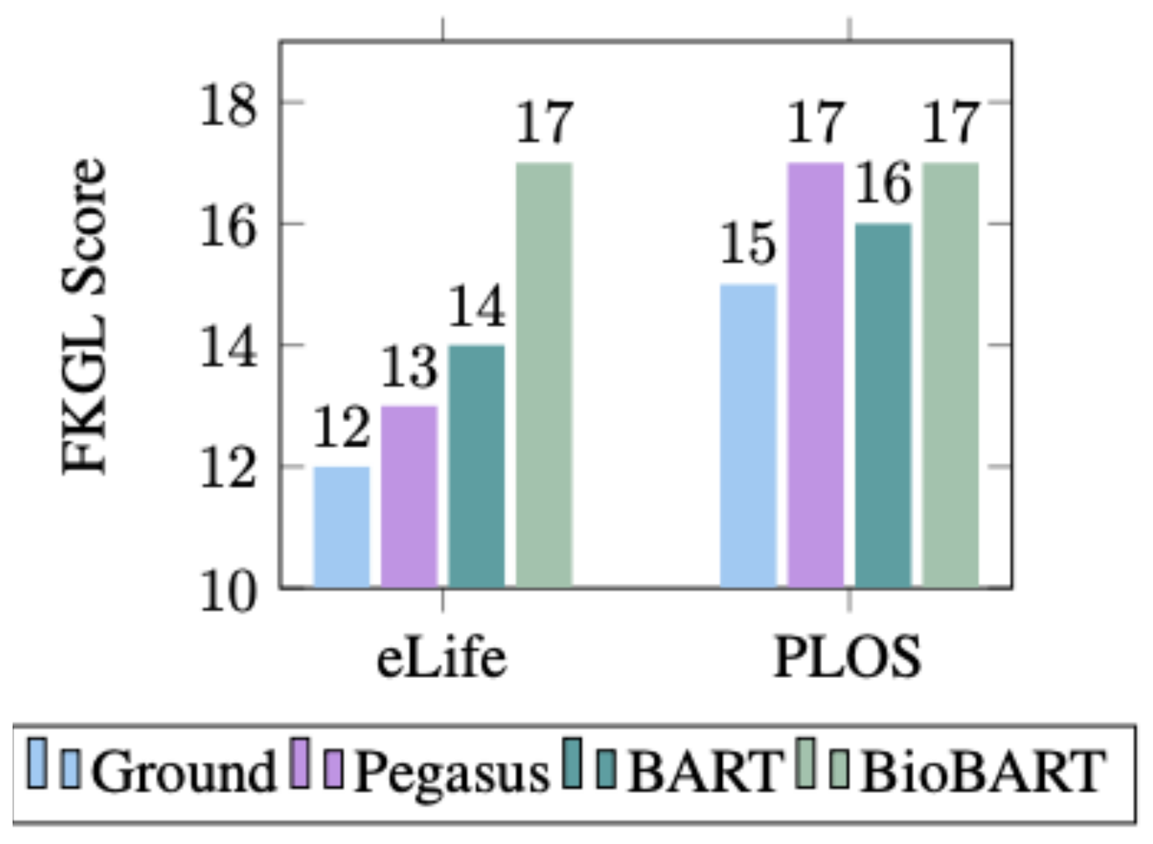


Fig 3. FKGL Scores for eLife and PLOS.

	MEDQA-AnS (p)			MEDQA-AnS (s)		
	R-1	R-2	R-L	R-1	R-2	R-L
TextRank	29.88	10.23	17.01	43.77	26.80	30.52
BART	**24.56**	**7.56**	**17.18**	**32.32**	15.42	**24.03**
Pegasus	17.44	5.36	13.44	19.54	7.46	14.93
PRIMERA	16.66	4.89	12.68	21.78	9.77	16.85
BioBART	23.16	7.47	16.47	30.87	**15.91**	23.66

Tab 2. Evaluation of Summarization Tasks (Multi-documents).

	PubMed			MIMIC-CXR			MEDQA-AnS (p)			MEDQA-AnS (s)		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Pegasus	45.97	20.15	28.25	22.49	11.57	20.35	18.29	4.82	13.87	22.21	8.23	16.76
BigBird	46.32	**20.65**	**42.33**	38.99	29.52	38.59	13.18	2.14	10.04	14.89	3.13	11.15
BART	44.16	20.28	36.80	**41.70**	**32.93**	**41.16**	**24.02**	**7.20**	**17.09**	38.19	22.20	30.58
SciFive	**48.83**	15.81	37.06	35.41	26.48	35.07	13.08	2.15	10.10	16.88	6.47	14.42
BioBART	-	-	-	41.61	32.90	41.00	22.58	7.49	16.69	**39.40**	**24.64**	**32.07**

Tab 3. Evaluation of Summarization Tasks (Single-document).

	eLife			PLOS		
	R-1	R-2	R-L	R-1	R-2	R-L
Pegasus	14.00	3.42	9.16	18.92	4.79	12.54
BART	**16.16**	**4.31**	**10.19**	21.09	7.20	14.17
BioBART	14.31	3.70	9.36	**23.80**	**7.83**	**15.65**

Tab 6. Lay Summarization Task Evaluation.

Dataset	HEAD-QA	MedMCQA
BioBERT	29.83	-
ClinicalBERT	29.43	-
BioMegatron	33.45	-
GatorTron	38.75	-
PubMedBERT	**42.52**	-

Tab 4. Evaluation of Multi-choice QA.

Dataset	LiveQA		
	R-1	R-2	R-L
Baize-healthcare	**20.47**	**5.03**	**18.73**
OPT-MedQuAD	17.04	3.93	15.94

Tab 5. Evaluation of Answer Generation.

Conclusion

- ***EHRKit:*** *We propose a Python library for clinical texts, including general API functions, MIMIC-specific functions and generative task-related functions.*
- ***Extend to more tasks and datasets:*** *We will test more medical datasets to provide benchmarks based on different NLP tasks in the medical field.*