Knowledge-Enhanced Few-Shot In-Context Learning for Medical Error Detection and Correction

Swati Rajwal[1], Eugene Agichtein[1], Abeed Sarker[2]
[1]Department of Computer Science, Emory University
[2]Department of Biomedical Informatics, Emory University

SCAN ME

INTRODUCTION

Medical Error Detection

Medical Error Sentence Detection

Medical Error Correction

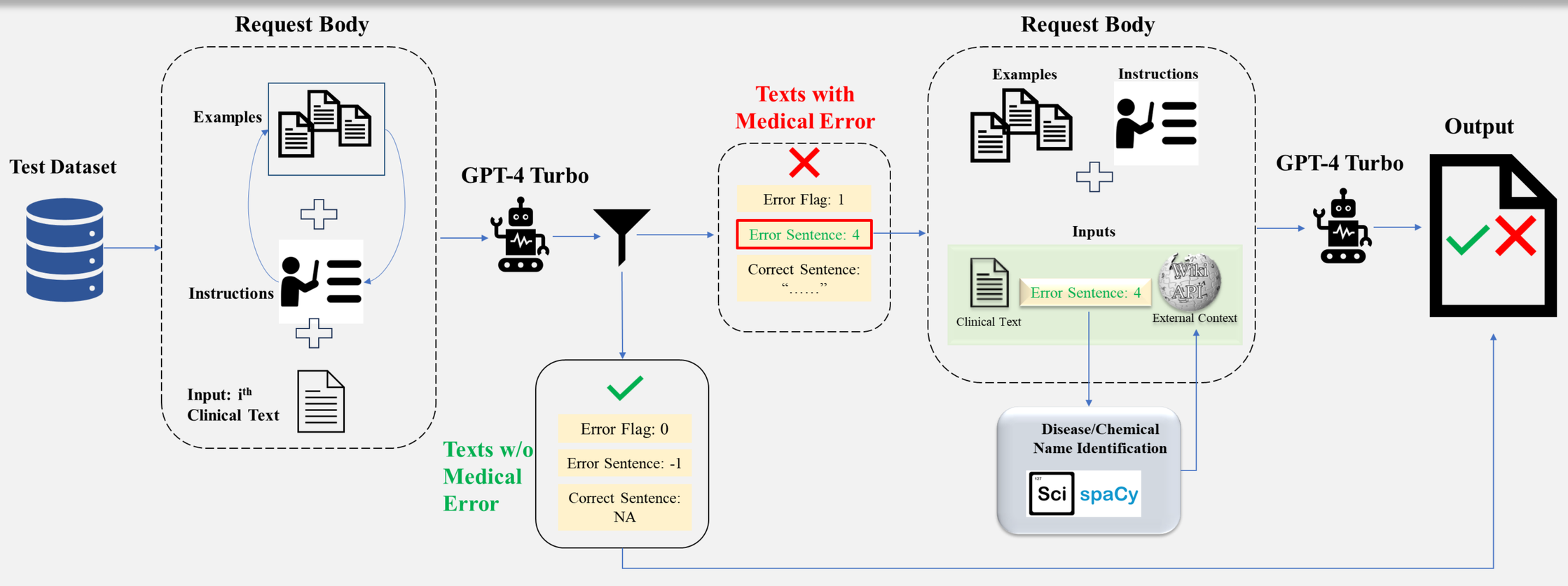
- Dataset by **Microsoft** and University of **Washington**
- Average length of clinical text: **781** words
- Error Flag: **0's** (no error), **1's** (medical error in text)
- No of observations: **3848** clinical texts

TASK RESULTS

Metrics	Validation		Test	
	GPT-4	RAG + GPT-4	GPT-4	RAG + GPT-4
Error Flag Accuracy	0.622	0.648	0.626	**0.68**[a]
Error Sentence Detection Accuracy	0.598	0.638	0.562	**0.64**[b]
Avg. Composite Score (NLG)	0.541	0.592	0.565	0.587

[a] Fourth,
[b] Second best accuracy among 17 participating teams in shared task

METHODOLOGY



MORE RESULTS*

Metric	FLAN T5	Mixtral	GPT-4	RAG + GPT4	Majority Voting
Precision	0.640	0.588	0.606	**0.767**	0.725
Recall	0.530	0.564	**0.884**	0.527	0.561
F_1 Score	0.580	0.576	**0.719**	0.625	0.633
Accuracy	0.573	0.538	0.617	**0.648**	0.638

* Experiments performed after shared task ended. Results only on validation dataset (n=574)

CONCLUSION

- Knowledge-enhanced few-shot learning promising for medical error detection & correction
- GPT4 struggled with rare/complex medical conditions
- ROUGE, BERTScore & BLEURT may not align with human judgment; expert evaluation needed