MEDSPEAK: A KNOWLEDGE GRAPH-AIDED ASR ERROR CORRECTION FRAMEWORK FOR SPOKEN MEDICAL QA

Yutong Song, Shiva Shrestha, Chenhan Lyu, Elahe Khatibi, Pengfei Zhang, Honghui Xu, Nikil Dutt, Amir Rahmani
University of California, Irvine, Kennesaw State University

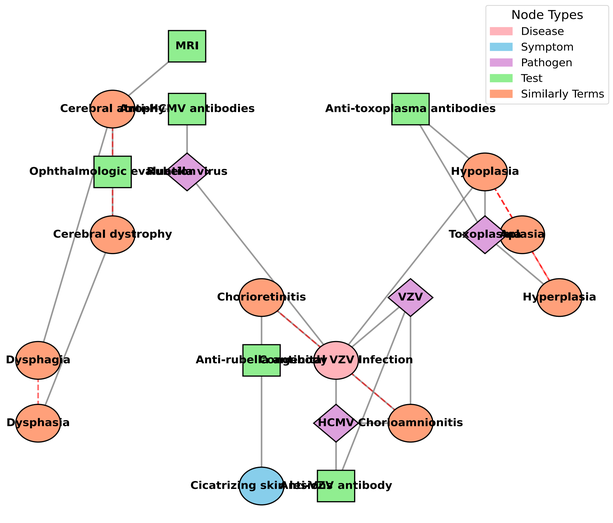
INTRODUCTION

Spoken question-answering (SQA) systems relying on automatic speech recognition (ASR) often struggle with accurately recognizing medical terminology.

Index Terms—Knowledge Graph, Automatic Speech Recognition, Spoken Question Answering, Large Language Models

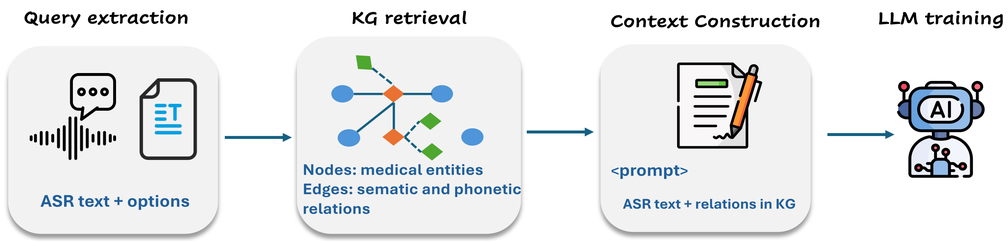
Example

[SPEECH]
A neonate presented with cicatrizing skin lesions all over the body with hypoplasia of all limbs. An MRI of the brain revealed diffuse cerebral atrophy. An ophthalmologic evaluation reveals chorioretinitis. Which of these tests is most likely to show a positive result in this patient?
[TEXT]
Option A: Anti-HCMV antibodies Option B: Anti-toxoplasma antibodies
Option C: Anti-VZV antibody Option D: Anti-rubella antibody
[OUTPUT] The correct answer is option **C**



We construct our knowledge graph using the Unified Medical Language System (UMLS)[1] dataset from the National Institutes of Health (NIH) , establishing relationships between medical terms through the MRREL table.

METHODOLOGY



Assistant Target is defined as a structured two-line completion format:
Corrected Text: [Original GT Text]
Correct Option: [Option Letter<A|B|C|D>]

Training Objective:
Given an input–output pair (x, y), the model is optimized with a single causal language modeling objective. Formally, the input sequence is defined as:

$$x = [s\,;\, u(ASR, Options, KG)]$$

The model parameters θ are trained to minimize:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{|y|} \log P_\theta(y_i \mid x, y_{<i})$$

This formulation unifies ASR error correction and multiple-choice reasoning into a single generation task. The KG ensures domain knowledge remains available, enabling the model to produce clinically accurate and context-aware outputs.

DATASETS

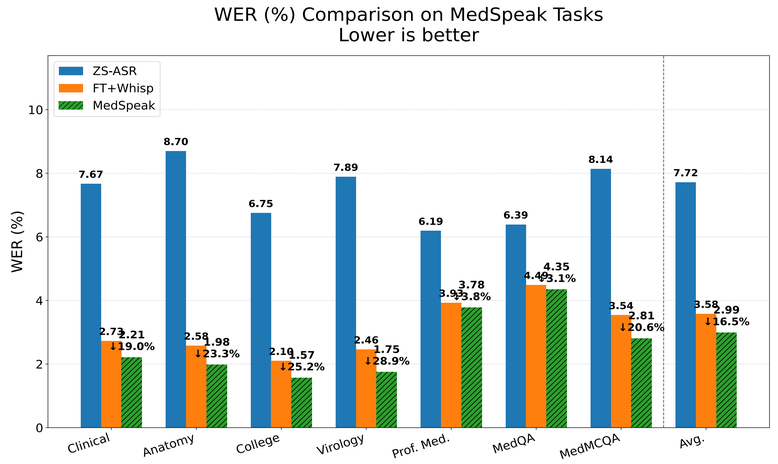
	Dataset	Dis.	Symp.	Diag. T.	Uniq. T.	Aud. T.
MMLU	Clinical Knowledge	677	59	45	677	5.2
	Anatomy	484	37	29	484	4.5
	College Medicine	344	29	22	344	6.0
	College Biology	250	21	18	250	3.9
	Medical Genetics	239	17	12	239	4.3
	Professional Medicine	1,145	113	95	1,145	8.7
Med QA (Test)		4,390	1,120	985	4,390	7.8
MedMCQA (Test)		19,978	2,319	1,983	11,978	10.5
MedMCQA (Train)		45,444	4,128	3,724	45,444	36.2

Each dataset is synthesized using OpenAI's TTS API (16,000 Hz, WAV mono format) with six distinct voices to enhance speaker variability and phonetic diversity [2].

RESULTS

Table 1. QA Accuracy of MedSpeak and Baselines.

Task	ZS-ASR	Zero-Shot	FT+Whisp	FT-LLM	MedSpeak
MMLU					
Clinical	62.2	66.3	85.6	94.3	95.4
Anatomy	57.0	64.4	85.2	94.1	93.3
College	62.1	67.5	84.2	92.7	95.6
Virology	47.6	48.8	85.5	91.0	95.8
Prof. Med.	74.6	76.1	88.6	94.9	97.8
MedQA	54.9	58.9	86.6	918	97.5
MedMCQA	45.5	52.8	82.3	92.5	91.5
Avg.	**50.2**	**56.3**	**83.7**	**92.5**	**93.4**



CONCLUSION

Our contributions include:
(1) An automated medical knowledge graph (KG) construction method encodes semantic and phonetic relationships between terms, enabling robust
error correction and reasoning.
(2) A multi-constraint retrieval mechanism leverages phonetic and semantic features to generate accurate ASR hypotheses for medical terminology.
(3) A fine-tuned LLM integrates retrieved knowledge with answer constraints, producing precise transcriptions and reliable answers.

CONTACT

yutons12@uci.edu
https://github.com/RainieLLM/MedSpeak

References:
[1] National Library of Medicine, "Unified medical language system (umls)," National Institutes of Health, 2024.
[2] Yanis Labrak, Adel Moumen, Richard Dufour, and Mickael Rouvier, "Zero-shot end-to-end spoken ques tion answering in medical domain," arXiv preprint, 2025.