Nevin M. Matasyoh*, Ramy A. Zeineldin, and Franziska Mathis-Ullrich

Optimising speech recognition using LLMs: an application in the surgical domain

https://doi.org/10.1515/cdbme-2024-1012

Abstract: Automatic speech recognition (ASR), powered by deep learning techniques, is crucial for enhancing human-computer interaction. However, its full potential remains unrealized in diverse real-world environments, with challenges such as dialects, accents, and domain-specific jargon, particularly in fields like surgery, persisting. Here, we investigate the potential of large language models (LLMs) as error correction modules for ASR. We leverage Whisper-medium or ASR-LibriSpeech for speech recognition, and GPT-3.5 or GPT-4 for error correction. We employ various prompting methods, from zero-shot to few-shot with leading questions and sample medical terms to correct wrong transcriptions. Results, measured by word error rate (WER), reveal Whisper's superior transcription accuracy over ASR-LibriSpeech, with a WER of 11.93% compared to 32.09%. GPT-3.5, with the few-shot with medical terms prompting method, further enhances performance, achieving a 64.29% and 37.83% WER-reduction for Whisper and ASR-LibriSpeech, respectively. Additionally, Whisper exhibits faster execution speed. Substituting GPT-3.5 with GPT-4 further enhances transcription accuracy. Despite some few challenges, our approach demonstrates the potential of leveraging domain-specific knowledge through LLM prompting for accurate transcription, particularly in sophisticated domains like surgery.

Keywords: Error correction, Zero-shot prompting, Few-shot-prompting, Large language model, Automatic speech recognition

1 Introduction

Automatic speech recognition (ASR) has experienced tremendous growth with the advent of deep learning techniques, becoming essential in human-machine interaction [1]. It has significantly evolved from simple or small-sized vocabulary applications to large-sized vocabulary applications, and from speaker-dependent to speaker-independent systems. This is evidenced by development of models such as the Amharic spoken digits recognition (AmSDR) [2], the Conformer [3], ASR-LibriSpeech [4] and the whisper-medium [5], among others. Additionally, ASR's scope has broadened from quiet to noisy environments, with error correction mechanisms being incorporated into ASR models [6] to enhance performance.

Despite notable progress, ASR systems face challenges in real-world applications due to noise, dialects, accents, and specialized jargon [7]. Environmental noises and varying dialects introduce pronunciation differences, while complex jargon and medical terminology pose interpretation difficulties, reducing system accuracy. To address these issues, newer techniques for automatic error detection and correction have been developed, shifting away from human intervention methods [8, 9] towards Language Models (LMs) as solutions for enhancing ASR system accuracy and preventing error propagation into subsequent modules [10, 11].

Language models, particularly advanced variants such as Large Language Models (LLMs), have demonstrated significant power in language understanding and generation, leading to a surge in research exploring their potential across various domains. Recent studies have specifically delved into utilizing LLMs for generative error correction (GER) in ASR, representing a shift from traditional model rescoring to considering multiple candidate hypotheses for predicting transcriptions [12, 13]. These studies focus on addressing issues such as code-switching [14] and denoising speech [6]. They achieve this by fine-tuning LLMs to map N-best hypotheses to ground-truth transcription, thereby improving the accuracy of transcriptions. Other studies explore the generative capabilities of LLMs through innovative prompting schemes without fine-tuning [13], demonstrating significant improvements even with limited training data.

Inspired by these successes, this study explores the potential of LLMs in error correction within speech recognition systems, with a specific focus on the surgical domain. We employ a prompting mechanism without fine-tuning for error correction. Our contributions can be summarized as follows:

– Presenting few-shot and zero-shot prompting mechanism for error correction in ASR systems, reducing computational complexity.

***Corresponding author: Nevin M. Matasyoh,**
Friedrich-Alexander-University Erlangen-Nürnberg, Department of Artificial Intelligence in Biomedical Engineering,
Werner-von-Siemens-Strasse 61, 91052 Erlangen, Germany,
e-mail: nevin.matasyoh@fau.de
Ramy A. Zeineldin, Franziska Mathis-Ullrich, Friedrich-Alexander-University Erlangen-Nürnberg, Department of Artificial Intelligence in Biomedical Engineering, Erlangen, Germany

– Demonstrating LLMs' effectiveness in addressing surgical-specific challenges, such as complex medical terminology.
– Evaluating our approach on a dataset specifically curated for surgical speech recognition, highlighting its applicability and effectiveness in real-world settings.

2 Methods

2.1 Models

ASR: We employ two state-of-the-art models for transcription: Whisper-medium [5], a transformer-based model trained on a diverse audio dataset for various speech tasks, and ASR-LibriSpeech [4], a conformer-based model designed for end-to-end speech recognition, pretrained on LibriSpeech data from SpeechBrain. These models are known for their exceptional performance and are well-suited for our task.
Error Correction: For transcription error correction, we utilize two LLMs, GPT-3.5 and GPT-4 [15], developed by OpenAI. These models exhibit advanced language understanding and generation capabilities, crucial for precise error correction.

2.2 Prompting techniques

Few-Shot: This approach entails providing instructions alongside leading questions to clarify the task to be performed. These questions are specifically related to the task or tailored to the domain, such as cerebral ventricular anatomy or intracranial structures in our experiments. Some of the questions include, *"Do you specifically understand the cerebral ventricular anatomy?"*, *"Can you explain error correction in automatic speech recognition systems?"*, and *"Do you understand the internal human anatomy?"*. The prompt may also incorporate sample medical terms to assist in correcting inaccurately transcribed terms. These terms are sample names of the intracranial structures that could be misspelled by the ASR system. By incorporating these specific terms within the prompting mechanism, we provide targeted guidance to the LLMs, facilitating more precise error correction. This approach acknowledges the potential limitations of relying solely on the models' intrinsic knowledge and seeks to augment their performance by supplying relevant contextual cues.
Zero-Shot: In contrast to few-shot prompting, zero-shot prompting involves instructing the LLMs to correct transcription errors without any specific leading questions or sample medical terms. This method capitalizes on the inherent generative power and knowledge within the LLMs to enhance transcription accuracy. By solely relying on the inherent capabilities of the models, we aim to assess their proficiency in error correction without external assistance.

3 Experiments

Dataset: We conduct experiments on our self-curated audio-transcription dataset specifically tailored to intracranial structures. We use gTTS(Google Text-to-Speech) [16] to convert 80 single-sentence descriptions of 15 intracranial structures into audio. Each description contains an average of 9 words. To ensure variability, we augment the audio descriptions into 7 different English dialects or accents. The resulting audio-transcription pairs are then used to evaluate the performance of a proposed method.
Implementation: First, an audio description is transcribed using whisper-medium or ASR-LibriSpeech models. The resulting transcription is then subjected to error correction using GPT-3.5 or GPT-4, employing both zero-shot and few-shot prompting techniques as described above.
Evaluation: We evaluate our method using Word Error Rate (WER), a metric commonly used to evaluate ASR systems [6, 13, 14]. WER measures the rate of errors between the reference transcription and the ASR-generated transcription [17]. WER is expressed as:

$$WER = \frac{S + D + I}{N}$$

where S is the number of substitutions, D is the number of deletions, and I is the number of insertions in the ASR output that should be corrected to transform the ASR output into the reference transcription. N is the total number of words in the reference transcription.

4 Results and Discussion

We evaluate the proposed method on our self-curated dataset, as described in Section 3. We utilize two ASR models, namely Whisper and ASR-LibriSpeech, along with two LLMs, GPT-3.5 and GPT-4, to ensure the general effectiveness of the method. Table 1 presents the results, using GPT 3.5, based on different prompting methods: zero-shot (ZS), few-shot with leading questions (FS_{Qns}), few-shot with sample medical terms (FS_{medTms}), and few-shot with both sample medical terms and leading questions ($\text{FS}_{Qns\ \&\ medTms}$). We present results in terms of WER of the transcribed sentences and speed of the method before error correction (WoC) and after error correction using LLM (WC).

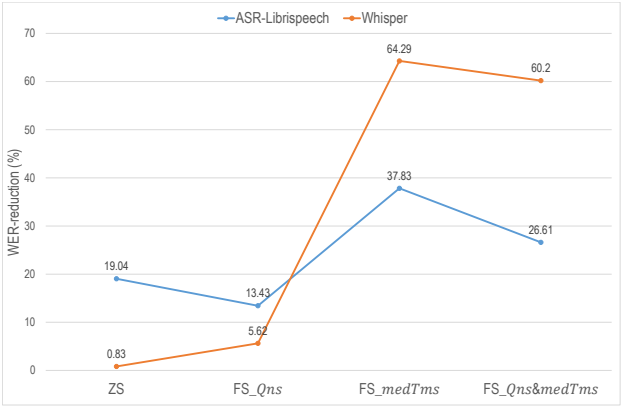

Fig. 1: Performance comparison in terms of WER-reduction (%) using GPT-3.5 with different prompting methods (ZS = zero-shot and FS = few-shot). Notations: (WoC) = without correction, (WC) = with correction, (Qns) = with leading questions, and ($medTms$) = with sample medical terms.

From the results, it can be seen that the Whisper model consistently outperforms ASR-LibriSpeech across all prompting methods. Without error correction, Whisper achieves very low WERs compared to LibriSpeech. At all four prompting methods, Whisper attains WERs between 11% and 12%, while ASR-LibriSpeech attains a WER of 32.09%, values which are approximately three times those of Whisper.

The performance of Whisper is further magnified by the use of LLM for error correction. Notably, when prompted with sample medical terms, we observed a substantial decrease in Whisper's WER from 11.93% to 4.26%, representing approximately a 64% WER reduction (refer to Figure 1). Similar results are observed when both medical terms and leading questions are included in the prompt. Whisper also outperforms ASR-LibriSpeech in terms of speed. This suggests that the inherent capabilities of Whisper, coupled with the error correction provided by LLMs, contribute significantly to its great performance.

Generally, few-shot prompting with medical terms (FS_{medTms}) outperforms all other prompting methods in both ASR models (Whisper and ASR-LibriSpeech). This is evident in Figure 1, where Whisper and ASR-LibriSpeech achieve the highest WER reduction of 64.29% and 37.83%, respectively, when prompted with sample medical terms. The substantial reduction in WER achieved by Whisper, augmented with LLM, particularly when prompted with medical terms, highlights the potential of leveraging domain-specific knowledge for more accurate transcription, especially in specialized fields such as surgical field.

We further assess the overall performance by evaluating the impact of a more advanced LLM version, GPT-4. To do so, we employ the most effective prompting method (FS_{medTms}) identified in Table 1. Figure 2 illustrates a comparison between the two ASR models (Whisper and ASR-LibriSpeech) when using both GPT-3.5 and GPT-4, focusing on WER reduction. There is notable improvement upon replacing GPT-3.5 with GPT-4. The WER reduction increases from 37.83% to 47.12% for ASR-LibriSpeech and from 64.29% to 67.66% for Whisper. This indicates that GPT-4 exhibits better generalization ability compared to GPT-3.5.

Tab. 1: Performance comparison in terms of WER (%) and speed (s) using different prompting methods (ZS = zero-shot and FS = few-shot), with error correction performed using GPT-3.5. Notations: (WoC) = without correction, (WC) = with correction, (Qns) = with leading questions, and ($medTms$) = with sample medical terms.

Prompting methods	Model	WER_{WoC}	WER_{WC}	$Time_{WoC}$	$Time_{WC}$
ZS	asr-librispeech	32.09	25.98	0.653	1.460
	whisper	12.01	11.91	0.426	1.093
FS_{Qns}	asr-librispeech	32.09	27.78	0.652	1.448
	whisper	11.93	11.26	0.421	1.076
FS_{medTms}	asr-librispeech	32.09	19.95	0.648	1.438
	whisper	11.93	4.26	0.420	1.111
$FS_{Qns\ \&\ medTms}$	asr-librispeech	32.09	23.55	0.650	1.448
	whisper	12.01	4.78	0.421	1.251

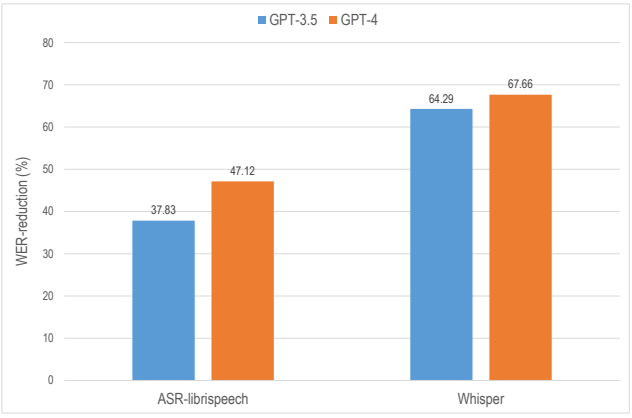

Fig. 2: Impact of different versions of LLMs (GPT-3.5 and GPT-4) on the performance of Whisper and ASR-Librispeech in terms of WER-reduction (%), using few-shot prompting with medical terms (FS_{medTms}).

Despite its impressive performance, our method exhibits some limitations. We highlight sample failures observed in different ASR models augmented by GPT-3.5, employing various prompting methods (refer to Table 2). Diverse accents or dialects can impact transcription accuracy. For instance, 'recessus' is transcribed as 'recess' or 'recesses'. Similarly, incorrect renditions of 'tuber cinereum' as 'tuber scenario', 'tubiscinarian', or 'tube of scenarium' indicate gaps in domain-specific knowledge within the LLM.

The utilization of gTTS can also introduce limitations, particularly in pronouncing complex medical terms, which

Tab. 2: Examples of incorrect transcriptions generated using GPT-3.5 with various prompting methods, along with the corresponding WER (%) compared to the ground truth.

Method	Transcription	WER
Ground Truth	An oval dark recessus next to the tuber cinereum	-
whisper + LLM	And oval dark recesses next to the tuber cinereum	22.22
	An oval dark recess next to the tuber cinereum	11.11
	An oval dark recessus next to the tuber scenario	11.11
	An oval dark recess next to the tubus scenario	33.33
asr-librispeech + LLM	An awful dark recesses next to the tubiscinarian	44.44
	An oval dark recesses next to the tuber sinuarum	22.22
	An awful dark recesses next to the tubiscinarian	44.44
	An oval dark recesses next to the tube of scenarium	44.44

may impact ASR system error rates due to mispronunciations. To enhance the authenticity and applicability of our findings, future research should utilize authentic audio recordings from surgical environments rather than synthesized speech. This would provide a dataset rich in natural speech variations and ambient operating room noises, enabling a more robust evaluation of ASR systems in real-world clinical settings.

In sensitive environments such as surgical rooms, privacy and data security are crucial. In such cases, alternative methods, such as using open-source LLMs or OpenAI's API with Enterprise-grade security features including zero data retention and encryption, can ensure compliance with strict privacy regulations.

5 Conclusion

We investigate the potential of LLMs in enhancing transcription accuracy in ASR. In our experiments, we utilize the Whisper and ASR-Librispeech models for speech recognition and employ GPT-3.5 or GPT-4 for error correction. Our findings suggest that through effective prompting, LLMs have the potential to significantly enhance the performance of ASR systems.

Author Statement

Research funding: The corresponding author is funded by the German Academic Exchange Service (DAAD) under Scholarship No. 91805762. Conflict of interest: Authors state no conflict of interest.

References

[1] Malik M, Malik MK, Mehmood K, Makhdoom I. Automatic speech recognition: a survey. Multimedia Tools and Applications. 2021;80:9411-57.

[2] Ayall TA, Zhou C, Liu H, Brhanemeskel GM, Abate ST, Adjeisah M. Amharic spoken digits recognition using convolutional neural network. Journal of Big Data. 2024;11(1):1-23.

[3] Gulati A, Qin J, Chiu CC, Parmar N, Zhang Y, Yu J, et al. Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:200508100. 2020.

[4] Ravanelli M, Parcollet T, Plantinga P, Rouhe A, Cornell S, Lugosch L, et al.. SpeechBrain: A General-Purpose Speech Toolkit; 2021. ArXiv:2106.04624.

[5] Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision. arXiv; 2022. Available from: https://arxiv.org/abs/2212.04356.

[6] Hu Y, Chen C, Yang CHH, Li R, Zhang C, Chen PY, et al. Large Language Models are Efficient Learners of Noise-Robust Speech Recognition. arXiv preprint arXiv:240110446. 2024.

[7] Lu X, Li S, Fujimoto M. In: Kidawara Y, Sumita E, Kawai H, editors. Automatic Speech Recognition. Singapore: Springer Singapore; 2020. p. 21-38. Available from: https://doi.org/10.1007/978-981-15-0595-9_2.

[8] Shi Y, Zhou L. Supporting dictation speech recognition error correction: the impact of external information. Behaviour & Information Technology. 2011;30(6):761-74.

[9] Errattahi R, El Hannani A, Ouahmane H. Automatic Speech Recognition Errors Detection and Correction: A Review. Procedia Computer Science. 2018;128:32-7. 1st International Conference on Natural Language and Speech Processing. Available from: https://www.sciencedirect.com/science/article/pii/S1877050918302187.

[10] Shin J, Lee Y, Jung K. Effective sentence scoring method using bert for speech recognition. In: Asian Conference on Machine Learning. PMLR; 2019. p. 1081-93.

[11] Leng Y, Tan X, Wang R, Zhu L, Xu J, Liu W, et al. Fastcorrect 2: Fast error correction on multiple candidates for automatic speech recognition. arXiv preprint arXiv:210914420. 2021.

[12] Chen C, Hu Y, Yang CHH, Siniscalchi SM, Chen PY, Chng ES. Hyporadise: An open baseline for generative speech recognition with large language models. Advances in Neural Information Processing Systems. 2024;36.

[13] Huck Yang CH, Gu Y, Liu YC, Ghosh S, Bulyko I, Stolcke A. Generative Speech Recognition Error Correction with Large Language Models and Task-Activating Prompting. arXiv e-prints. 2023:arXiv-2309.

[14] Chen C, Hu Y, Yang CHH, Liu H, Siniscalchi SM, Chng ES. Generative error correction for code-switching speech recognition using large language models. arXiv preprint arXiv:231013013. 2023.

[15] OpenAI. GPT-3: Generative Pre-trained Transformer 3; 2024. Accessed on: 2024-05-15. Available at: https://openai.com/api/.

[16] Durette PN. gTTS: Google Text-to-Speech; 2024. Available at: https://pypi.org/project/gTTS/.

[17] Wang Q, Zheng Z, Wang Q, Deng D, Zhang J. Generalizations of wearable device placements and sentences in sign language recognition with Transformer-based model. IEEE Transactions on Mobile Computing. 2024.