Zero-Shot End-To-End Spoken Question Answering In Medical Domain

Yanis Labrak, Adel Moumen, Richard Dufour, Mickaël Rouvier

Zero-Shot End-To-End Spoken Question Answering In Medical Domain

Yanis Labrak[1,2]*, Adel Moumen*[1]*, Richard Dufour*[1,3]*, Mickael Rouvier*[1]

[1]LIA - Avignon University, France [2]Zenidoc, France
[3]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

yanis.labrak@univ-avignon.fr, adel.moumen@univ-avignon.fr,
richard.dufour@univ-nantes.fr, mickael.rouvier@univ-avignon.fr

Abstract

In the rapidly evolving landscape of spoken question-answering (SQA), the integration of large language models (LLMs) has emerged as a transformative development. Conventional approaches often entail the use of separate models for question audio transcription and answer selection, resulting in significant resource utilization and error accumulation. To tackle these challenges, we explore the effectiveness of end-to-end (E2E) methodologies for SQA in the medical domain. Our study introduces a novel zero-shot SQA approach, compared to traditional cascade systems. Through a comprehensive evaluation conducted on a new open benchmark of 8 medical tasks and 48 hours of synthetic audio, we demonstrate that our approach requires up to 14.7 times fewer resources than a combined 1.3B parameters LLM with a 1.55B parameters ASR model while improving average accuracy by 0.5%. These findings underscore the potential of E2E methodologies for SQA in resource-constrained contexts.

Index Terms: spoken question answering, large language model, medical, zero-shot, whisper, ssl

1. Introduction

Spoken Question Answering (SQA) aims to identify the correct answer from spoken documents or texts in response to a given spoken query. Unlike many other spoken language understanding tasks, such as speech summarization, which primarily focus on semantic comprehension at the utterance level, SQA demands advanced comprehension and reasoning over extensive audio content. In addition to grasping the question and understanding the global context within the audio, it requires capturing nuanced details to accurately select the correct answer, often involving the utilization of out-of-context information. As a result, SQA poses a significant challenge due to its multifaceted nature.

Traditionally, SQA methods comprise a cascade of an Automatic Speech Recognition (ASR) system to transcribe the audio question followed by a Language Model (LM). The LM takes a prompt and the automatic transcription as input to predict the correct answer among a list of options. However, ASR errors introduce noise into the LM input, leading to performance degradation and information loss, despite community efforts [1, 2, 3] to mitigate these issues and enhance robustness to transcription errors. Consequently, the cascade of stages cannot match the performance of a single-stage model based on speech due to inherent information loss [4].

The emergence of Large Language Models (LLMs) like Bloom [5] or LLaMa 2 [6] represents a significant advancement in question-answering systems. However, these models require extensive parameter scaling, further complicating the challenge of running separate models for each stage. For instance, the ASR models are already large (e.g., Whisper Medium with 769M parameters and Large V2 with 1.55B parameters), necessitating significant hardware resources for each stage. Consequently, there is a growing interest in directly extracting information from speech to preserve maximal information while minimizing hardware requirements.

Several architectures, such as Whisper [7], CLAP [8], and SpeechT5 [9], have proposed unifying textual and audio modalities using encoder-decoder models. Notably, autoregressive approaches based on LLMs, exemplified by SpeechGPT [10], have emerged. These models rely on textual prompts to encode speech signals into discrete units.

We propose a novel end-to-end audio-text entailment strategy for zero-shot multiple-choice question answering tasks, focusing on the medical domain. Inspired by zero-shot classification methods in textual Natural Language Processing (NLP) [11, 12] and computer vision [13, 14], our approach leverages the model's capacity to identify modalities that entail each other. Our contributions include:

- An innovative audio-text entailment approach for zero-shot spoken multiple-choice question answering tasks.
- A new SQA dataset tailored to the medical domain.
- A zero-shot performance comparison of 4 existing state-of-the-art end-to-end models.
- An in-depth analysis of the disposition of the information required for the SQA task within speech encoder layers.
- A public release of all the code and data on GitHub and Hugging Face [1].

2. Medical spoken question answering

In this section, we define the SQA task (Section 2.1) and present the open benchmark constructed from established medical datasets initially in textual format (Section 2.2). Additionally, we describe the audio prompt format (Section 2.3) and the SQA evaluation protocol (Section 2.4).

2.1. Definition

We focus on multiple-choice SQA within the medical domain. Each instance comprises an audio question followed by four possible spoken responses, denoted as (q, o, c, a). Here, q represents the question, o denotes the options (labeled A to D), c indicates the correct answer and a encapsulates the audio containing both the question and options. Questions are structured

[1]https://huggingface.co/SpokenMedicalQA

as single-turn interactions, devoid of dialogue. This evaluation relies solely on the model's internal knowledge without external information or span extraction. The primary objective is to assess end-to-end model performance in understanding and accurately choosing the correct answer from spoken input.

2.2. Tasks collection and description

Recent years have seen significant progress in SQA datasets, such as Clotho-AQA [15], Spoken-SQuAD [16], and LibriSQA [17]. However, these datasets do not specifically target the healthcare domain or rely solely on audio inputs. The absence of SQA datasets in the medical domain hampers the development of question answering systems tailored to healthcare contexts. To address this gap, we propose synthesizing an audio dataset from existing textual multiple-choice question answering (MCQA) corpora. Our approach involves using Text-To-Speech (TTS) technology on these MCQA textual datasets to generate synthetic audios, leveraging advancements in TTS models that increasingly resemble human speech quality [18, 19]. We utilized the OpenAI TTS API (`tts-1`) to synthesize speech based on the questions and available options. The speakers were alternated through the 6 available voices to introduce diversity and realism into the dataset. The resulting audio files were sampled at 16,000 Hz and converted to WAV mono channel format.

Our reference texts were sourced from three open-source textual MCQA corpora in English, all relevant to healthcare, featuring single possible answers and a four-option format. Note that only the test data are detailed here, as the proposed approaches operate under zero-shot conditions.

MMLU [20] comprises exam questions spanning 57 subjects, including those relevant to healthcare. We focused on six healthcare-related subjects already evaluated in MedPaLM-2 [21]: college biology, college medicine, anatomy, professional medicine, medical genetics, and clinical knowledge. The dataset includes a test set of 1,089 questions, totaling 8 hours and 39 minutes of synthesized audio.

MedQA [22] integrates questions formatted similarly to the US Medical License Exam (USMLE), covering diverse medical topics. We exclusively utilized the test set, comprising 1,273 questions amounting to 21 hours and 22 minutes of audio.

MedMCQA [23] consists of questions with four options each, extracted from Indian medical entrance examinations (AIIMS/NEET). It encompasses 2,400 healthcare topics across 21 medical subjects, with 4,183 questions for the validation which are used as test ones since it is unavailable to the public [24]. The test set comprises 17 hours and 40 minutes of audio.

Our final benchmark encompasses 8 SQA tasks (including 6 from MMLU) derived from these 3 synthesized datasets. Table 2 summarizes the audio duration distribution according to the different labels available in the test set.

Table 2: *Audio duration distribution according to the labels.*

	MMLU	MedQA	MedMCQA	Total	# Doc.
A	1h50	5h55	5h41	13h28	1,936
B	1h54	5h08	4h31	11h33	1,648
C	1h50	5h49	3h57	11h37	1,519
D	3h03	4h28	3h30	11h03	1,442
Total	8h39	21h22	17h40	47h41	6,545

2.3. Audio prompt format

We standardized all textual MCQA datasets and synthesized them into audio format. These audio MCQAs serve as prompts for the studied and proposed SQA systems. Following experimentation with various formats and careful listening to the resulting audio outputs, we identified an effective format exemplified below:

Prompt Format

A 39-year-old woman, with a history of thyroidectomy and primary hyperparathyroidism presents for surgical evaluation for a right adrenal mass. **Preoperatively, which of the following medications should she receive to prevent a hypertensive emergency intraoperatively?** Option A: Atenolol Option B: Labetolol Option C: Nifedipine **Option D: Phenoxybenzamine**
The correct answer is Option **D**

2.4. Evaluation metric

The evaluation of multi-choice SQA with a single correct answer resembles a multi-class classification task. The performance is here assessed for each task using *Accuracy*, which measures the proportion of correctly predicted answers compared to the total number of questions. A prediction is considered accurate if it exactly matches the ground-truth answer, otherwise, it is classified as incorrect. Choosing the accuracy enables direct comparison with previous works on textual datasets [25, 26].

3. Studied and proposed methods

This section outlines the zero-shot approaches studied for SQA. Firstly, we introduce baseline models with cascade systems (Section 3.1). Then, we present models integrating our end-to-end audio-text entailment approach (Section 3.2).

3.1. Baseline cascade approaches

Our baseline models involve a two-stage process: transcription of audio inputs into text using an ASR module, followed by its processing with an LLM to select the correct answer to posed questions. We conducted experiments with various models to assess the impact of different ASR and LLM configurations on SQA performance. In the ASR stage, we compared the performance using the reference transcription (*Oracle*) against Whisper Small, Medium, and Large V2 ASR models to identify potential transcription error propagation issues. Subsequently, in the LLM stage, we compared the performance of an LLM similar in size to Whisper Large V2 (1.5 billion parameters), named Phi 1.5, against larger models based on the LLaMa 2 architecture, configured with 7B and 13B parameters, to assess the scalability of performance with model size. In total, we investigated 12 cascade system combinations.

During the second step of inference, the LLM predicts the next token based on the input prompt, generating probabilities for each token in the vocabulary. To ensure relevance, the vocabulary is filtered to include only relevant tokens (in this case, choice letters) corresponding to the expected answer options. This approach prevents the model from generating irrelevant tokens or hallucinations [27].

Table 1: *Accuracy (in %) of the zero-shot cascade methods. Highest value in bold and second best is underlined.*

		MMLU								
		Clinical KG	**Medical Genetics**	**Anatomy**	**Pro Medicine**	**College Biology**	**College Medicine**	**MedQA**	**MedMCQA**	**Avg.**
Phi 1.5	Oracle	31.3	**39.0**	19.3	20.6	29.2	28.9	27.7	31.2	28.4
	Whisper Small	26.8	24.0	31.9	27.6	25.0	23.1	25.5	25.9	26.2
	Whisper Medium	27.9	20.0	35.6	27.6	25.7	24.9	25.4	25.4	26.6
	Whisper Large V2	31.7	19.0	34.1	24.6	26.4	26.0	27.6	26.2	27.0
Llama 2 7B	Oracle	21.5	30.0	18.5	18.4	25.7	20.8	27.7	32.1	24.3
	Whisper Small	29.4	31.0	25.2	33.5	31.9	31.2	29.9	30.7	30.3
	Whisper Medium	30.6	**39.0**	25.2	35.3	37.5	29.5	29.5	31.1	32.2
	Whisper Large V2	31.7	38.0	26.7	33.5	29.9	31.8	28.7	30.8	31.4
Llama 2 13B	Oracle	21.5	30.0	18.5	18.4	25.7	20.8	27.7	32.1	24.3
	Whisper Small	35.8	35.0	39.3	35.7	41.0	28.9	36.2	34.0	35.7
	Whisper Medium	**37.7**	36.0	**45.2**	**39.0**	**44.4**	**32.4**	**37.4**	**34.1**	**38.3**
	Whisper Large V2	34.7	38.0	37.0	**39.0**	39.6	**32.4**	36.8	33.1	36.3

3.2. Zero-shot end-to-end entailment-based approaches

Numerous studies [11, 28] have underscored the advantages of leveraging Natural Language Inference (NLI) for textual zero-shot entailment and classification tasks. However, except for CLAP [29] and Pengi [30], based on contrastive learning and prefix-tuning respectively, a limited adaptation of such methodologies has been observed in speech-related literature, particularly with large-scale pre-trained audio models like Whisper and SpeechGPT. Our proposed zero-shot audio-text entailment method is integrated into the four previously mentioned models, aiming to assess the likelihood of a textual sequence matching an audio recording. In our setup, the audio contains the question and options, while the text represents classes A to D.

For Whisper [7], we utilize audio features and request individual log probabilities for each letter using the format: <|*startoftranscript*|> *[A]* <|*endoftext*|>. The predicted class is determined by the highest average log probability. To comply with Whisper's 30-second limit for audio segments, we truncate segments beyond this duration to capture only the question and options. For SpeechGPT [10], we populate the model's context in a prompt filled with speech units obtained from HuBERT [31] representations discretized using k-means clustering on 1,000 clusters. We then request the generation of one additional token to the model. Subsequently, we filter the vocabulary to retain only the log probabilities corresponding to letters A to D, as described earlier in Section 3.1. Pengi [30] undergoes minimal changes in the model, audio representation, and prompt format, maintaining a similar procedure. The approach is slightly adapted for the CLAP model [29], a dual encoder architecture trained with contrastive language-audio pre-training. Here, individual encoders process both speech and text. Given an audio sample (a) and a list of classes (o), we identify the best match among all pairs by calculating the cosine distance between their vector representations. The pair with the closest distance is considered the predicted match.

4. Results

In this section, we examine the zero-shot condition performance on our SQA tasks using first the baseline cascade models (Section 4.1), and then our entailment approach across various end-to-end models (Section 4.2).

4.1. Zero-shot cascade approaches

Table 4 outlines the transcription performance, measured in Word Error Rate (WER), of Whisper ASR versions (Small, Medium, and Large V2) across various SQA tasks. Generally, Whisper Large V2 shows improved WER performance, except in MMLU Anatomy, where Whisper Medium performs better.

Table 4: *Transcription performance (in WER) on each SQA task. Best result in bold and second best is underlined.*

Tasks		**Whisper**		
		S	**M**	**L-V2**
MMLU	Clinical KG	5.45	4.21	**3.30**
	Medical Genetics	6.19	4.59	**4.31**
	Anatomy	4.90	**2.68**	3.50
	Pro Medicine	5.66	4.68	**4.54**
	College Biology	4.54	2.91	**2.66**
	College Medicine	26.02	25.54	**24.74**
	MedQA	7.50	6.21	**5.84**
	MedMCQA	7.99	6.33	**6.10**
Average		8.53	7.14	**6.87**

Table 1 displays the accuracy performance of studied LLM-based zero-shot cascade methods using Whisper automatic transcriptions on multiple SQA tasks. Interestingly, the Whisper model with the lowest WER might not always be the optimal choice in a cascade approach, indicating a lack of direct correlation between WER and SQA accuracy. Conversely, SQA performance appears to depend on LLM size, with larger models yielding higher accuracy. Notably, there is an 11.67% difference between Phi 1.5 and LLaMa 2 13B in Whisper Medium results, highlighting the significant advantage of scaling up LLMs. Except for Phi 1.5, all models show improved performance with transcriptions compared to Oracle. This enhancement, particularly in LLaMa 2 architectures, may be attributed to their better adaptability to speech normalization formats, reduced punctuation, and increased noise.

Furthermore, with LLaMa 2, Whisper Medium transcriptions emerge as the top performers. Notably, LLaMa 13B demonstrates a 1.95% overall accuracy gain over Whisper Large V2 and a 2.54% improvement over Whisper Small. Similar trends are observed in the 7B model, with increases of 0.8% over Large V2 and 1.9% over Small. The performance of the LLaMa 2 13B model in a zero-shot scenario with Whisper Medium transcriptions shows promising results.

4.2. Zero-shot end-to-end models' capabilities

Table 3 outlines the accuracy performance of zero-shot end-to-end models using our entailment method on our multiple-choice SQA benchmark. While the overall average accuracy remains similar across models, specific models demonstrate

Table 3: *Accuracy (in %) of the zero-shot end-to-end models applying our entailment method. Highest value in bold and second best is underlined, excluding SpeechGPT + Oracle (model aligned with reference transcriptions).*

		MMLU								
		Clinical KG	**Medical Genetics**	**Anatomy**	**Pro Medicine**	**College Biology**	**College Medicine**	**MedQA**	**MedMCQA**	**Avg.**
Whisper	Small	24.1	**31.0**	20.0	17.6	25.0	20.2	**27.7**	30.6	24.5
	Medium	**30.6**	20.0	17.8	42.6	26.4	30.6	21.9	22.5	26.5
	Large V2	27.5	24.0	26.7	20.2	20.1	19.6	25.8	27.4	23.9
CLAP	Unfused	26.8	23.0	24.4	37.1	**29.2**	32.9	23.1	19.7	27.0
	Large General	29.4	21.0	23.7	**44.5**	25.7	**34.1**	21.1	20.3	**27.5**
	Fused	21.5	30.0	18.5	18.4	25.7	20.8	**27.7**	**32.0**	24.3
Pengi	Base	24.9	26.0	**32.6**	21.3	19.4	24.8	24.0	24.4	24.7
	Base No Text Encoder	26.8	26.0	25.2	20.2	22.2	20.8	24.3	25.9	23.9
SpeechGPT	E2E	28.3	23.0	29.6	17.6	21.5	27.2	26.4	23.4	24.6
SpeechGPT	Oracle	36.2	32.0	27.4	35.7	29.9	34.1	24.4	27.2	30.8

proficiency in particular tasks, with none consistently outperforming others across all tasks. Notably, Whisper Medium showcases competitive zero-shot performance, surpassing cascade setups with Phi 1.5 despite having approximately half the parameters. CLAP's contrastive modeling outperforms Phi 1.5 but falls short of LLaMa 2 7B. Impressively, despite its smaller size—153M parameters in its base form and 193M in its larger form—CLAP performs remarkably well, being 14.7 times smaller than Whisper Large V2 combined with Phi 1.5 and 44.3 times smaller with LLaMa 2 7B. SpeechGPT encounters challenges in zero-shot tasks from speech, contrasting its performance with text (Oracle), highlighting difficulties in directly handling speech modality representations, which need to be addressed in the future, with a better alignment approach. Notably, Whisper, especially Whisper Medium, occasionally outperforms cascade configurations with Phi 1.5 in zero-shot scenarios. Specific tasks exhibit varying levels of difficulty for different models; for instance, MedMCQA yields high results with Whisper Small and CLAP Fused, while MMLU College Medicine favors Whisper Medium, CLAP Unfused, and CLAP Large General. SpeechGPT generally underperforms across most tasks, except for MMLU Anatomy and MedQA, where it outperforms most other models. Despite the small performance improvement over cascade systems, which is linked to the zero-shot setting, E2E systems can be enhanced by scaling with better quality SQA data and increasing the number of parameters to see if they follow scaling laws similar to LLMs.

5. Analysis of encoder layers

This section presents an extensive analysis to pinpoint the critical location of information crucial for SQA tasks within the layers encoding the audio signal. To conduct this analysis, we extracted a subset of the MedMCQA training set consisting solely of audio sequences shorter than 30 seconds, which comprised 97.56% of the data, resulting in 120 hours of spoken data. This subset was partitioned into training and validation sets using an 80%/20% ratio, yielding 95 hours and 23 hours, respectively. Our experimental approach involves fine-tuning audio encoders and introducing an intermediate trainable layer of equal size to the number of encoder layers. This intermediate layer selects information from the encoder's layers through a weighted sum of their representations when feeding the classification head. The objective of this weighted encoder layers approach is to analyze the necessity of specific layers for executing the SQA task while enhancing model understanding.

As depicted in Figure 1, illustrating cumulative weights across encoder layers, Whisper models exhibit a propensity to concentrate information in the final layers, aligning with prior research findings [32]. This indicates that these audio-based models effectively utilize the last layer to represent textual information, possibly due to heavy reliance on the decoder.

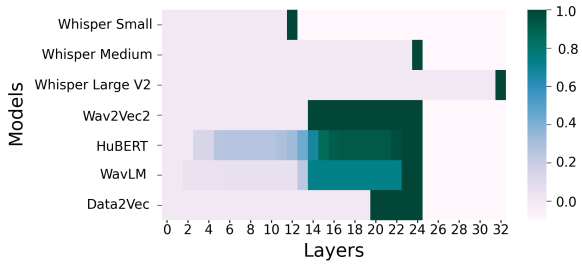

Figure 1: *Cumulative weights according to encoder layers.*

In contrast, Wav2Vec [33] and Data2Vec [34] primarily rely on a single intermediate layer, specifically the 15th and 21st layers, respectively. However, HuBERT [31] and WavLM [35] adopt a different strategy, integrating information from a broader range of layers. HuBERT integrates data from 12 layers, while WavLM incorporates information from 4 layers distributed across various regions of the encoder.

6. Conclusion

This study introduces a novel synthetic Spoken Question Answering (SQA) dataset tailored specifically to the medical domain. We conducted zero-shot comparative analyses of end-to-end speech methodologies using a new entailment technique against cascade speech transcription and LLM module. Our experiments and analysis demonstrate the effectiveness of our end-to-end approach, yielding performances comparable to those achieved by cascade models of similar sizes. Moving forward, we aim to explore the utilization of speech alignment techniques with LLMs to enhance end-to-end question answering performance, with a particular emphasis on improving outcomes in low-resource domains such as healthcare. Our research faced multiple constraints. Using limited speaker variety for synthetic audio may reduce accuracy compared to natural speech, affecting response precision. Simplifying task formulation lacks genuine human interaction dynamics but enables metric-based assessments, enhancing model reproducibility and cost efficiency. Finally, our study neglects multilingual contexts, highlighting the need for additional exploration in diverse linguistic settings.

7. Acknowledgements

This work was performed using HPC resources from GENCI-IDRIS (Grant 2024-AD011015344 and Grant 2022-AD011013061R2). This work was financially supported by ANR MALADES (ANR-23-IAS1-0005) and Zenidoc.

8. References

[1] C.-H. Lee, Y.-N. Chen, and H.-Y. Lee, "Mitigating the impact of speech recognition errors on spoken question answering by adversarial domain adaptation," in *ICASSP*, 2019, pp. 7300–7304.

[2] C.-H. Lee, S.-L. Wu, C.-L. Liu, and H. yi Lee, "Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension," in *Interspeech*, 2018, pp. 3459–3463.

[3] C. You, N. Chen, and Y. Zou, "Knowledge distillation for improved accuracy in spoken question answering," in *ICASSP*, 2021, pp. 7793–7797.

[4] H. Inaguma, K. Duh, T. Kawahara, and S. Watanabe, "Multilingual end-to-end speech translation," in *ASRU*, 2019, pp. 570–577.

[5] T. L. Scao, A. Fan, C. Akiki, and et al., "Bloom: A 176b-parameter open-access multilingual language model," 2023.

[6] H. Touvron, L. Martin, K. Stone, and et al., "Llama 2: Open foundation and fine-tuned chat models," 2023.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[8] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP*, 2023.

[9] J. Ao, R. Wang, L. Zhou, and et al., "SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing," in *ACL*, 2022, pp. 5723–5738.

[10] D. Zhang, S. Li, X. Zhang, and et al., "SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities," in *EMNLP*, 2023, pp. 15 757–15 773.

[11] K. Halder, A. Akbik, J. Krapac, and R. Vollgraf, "Task-aware representation of sentences for generic text classification," in *COLING*, 2020, pp. 3202–3213.

[12] M. Pàmies, J. Llop, F. Multari, N. Duran-Silva, C. Parra-Rojas, A. Gonzalez-Agirre, F. A. Massucci, and M. Villegas, "A weakly supervised textual entailment approach to zero-shot text classification," in *EACL*, 2023, pp. 286–296.

[13] Y. Du, J. Li, T. Tang, W. X. Zhao, and J.-R. Wen, "Zero-shot visual question answering with language model feedback," in *Findings of the ACL*, Toronto, Canada, 2023, pp. 9268–9281.

[14] O.-B. Mercea, L. Riesch, A. S. Koepke, and Z. Akata, "Audio-visual generalised zero-shot learning with cross-modal attention and language," in *CVPR*, June 2022, pp. 10 553–10 563.

[15] S. Lipping, P. Sudarsanam, K. Drossos, and T. Virtanen, "Clotho-aqa: A crowdsourced dataset for audio question answering," 2022.

[16] C.-H. Lee, S.-L. Wu, C.-L. Liu, and H.-y. Lee, "Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension," *Interspeech*, pp. 3459–3463, 2018.

[17] Z. Zhao, Y. Jiang, H. Liu, Y. Wang, and Y. Wang, "Librisqa: Advancing free-form and open-ended spoken question answering with a novel dataset and framework," 2023.

[18] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *ICML*, vol. 139, 2021, pp. 5530–5540.

[19] J. Kong, J. Kim, and J. Bae, "Hifi-gan: generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS*, 2020.

[20] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," in *ICLR*, 2021.

[21] K. Singhal, T. Tu, J. Gottweis, and et al., "Towards expert-level medical question answering with large language models," 2023.

[22] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," 2020.

[23] A. Pal, L. K. Umapathi, and M. Sankarasubbu, "Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering," in *Proceedings of the Conference on Health, Inference, and Learning*, vol. 174, 2022, pp. 248–260.

[24] C. Wu, W. Lin, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Pmc-llama: Towards building open-source language models for medicine," 2023.

[25] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems," 2023.

[26] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, and R. Dufour, "Biomistral: A collection of open-source pretrained large language models for medical domains," 2024.

[27] P. Liang, R. Bommasani, T. Lee, and et al., "Holistic evaluation of language models," *Transactions on Machine Learning Research*, 2023.

[28] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *ICML*, vol. 37, 2015, pp. 2152–2161.

[29] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP*, 2023, pp. 1–5.

[30] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An audio language model for audio tasks," in *NeurIPS*, 2023.

[31] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 3451–3460, 2021.

[32] H. Yang, J. Zhao, G. Haffari, and E. Shareghi, "Investigating Pretrained Audio Encoders in the Low-Resource Condition," in *Interspeech*, 2023, pp. 1498–1502.

[33] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[34] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *ICML*, vol. 162, 2022, pp. 1298–1312.

[35] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.