CORAAL QA: A DATASET AND FRAMEWORK FOR OPEN DOMAIN SPONTANEOUS SPEECH QUESTION ANSWERING FROM LONG AUDIO FILES

Natarajan Balaji Shankar[1] *, Alexander Johnson*[1]*, Christina Chance*[2]*, Hariram Veeramani*[1]*, Abeer Alwan*[1]

University of California, Los Angeles
[1]Department of Electrical and Computer Engineering, [2]Department of Computer Science

ABSTRACT

This paper presents a novel dataset (CORAAL QA) and framework for audio question-answering from long audio recordings containing spontaneous speech. The dataset introduced here provides sets of questions that can be factually answered from short spans of a long audio files (typically 30min to 1hr) from the Corpus of Regional African American Language. Using this dataset, we divide the audio recordings into 60 second segments, automatically transcribe each segment, and use PLDA scoring of BERT-based semantic embeddings to rank the relevance of ASR transcript segments in answering the target question. In order to improve this framework through data augmentation, we use large language models including ChatGPT and Llama 2 to automatically generate further training examples and show how prompt engineering can be optimized for this process. By creatively leveraging knowledge from large-language models, we achieve state-of-the-art question-answering performance in this information retrieval task.

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

1. INTRODUCTION

Recent advancements in BERT [1] and GPT [2]-based language models have revolutionized performance in question answering and information retrieval tasks on text. Now, a desirable outcome is to replicate the performance of these systems in the speech domain. That is, given a set of audio recordings and a user's input query for information, we seek to return audio recordings or spans that are relevant to the query. Successful architectures for this task typically take one of two frameworks: a cascade system or an end-to-end model. A cascade system first uses automatic speech recognition (ASR) to transcribe a spoken document and then passes that transcript to a downstream language model for text-based question answering. End-to-end systems seek to bypass the need for transcription and answer a question directly from audio features. Notable cascade models include [3] which introduces a self-supervised dialogue learning framework from conversational question answering and [4] which proposes a unified pipeline for multiple spoken language understanding tasks. End-to-end spoken question answering models of interest include SpeechBERT [5], which jointly encodes audio and text information for downstream spoken question answering, GhostT5 [6] which extracts and passes a lightweight speech feature representation to a pre-trained language model to answer questions from speech without the need for complete automatic speech recognition (ASR) transcription, and [7] which implements a dual attention mechanism for smoother incorporation of both text and audio. While end-to-end models show promise in eliminating errors propagated by ASR systems [8], cascade models are able to leverage large language models trained on massive amounts of text data for open domain question-answering. Currently, these cascade models may be especially preferable in low-resource applications for which there does not exist enough in-domain data to effectively train an end-to-end model from scratch. End-to-end systems may match or surpass the performance of cascade models as more labeled data for spoken question answering becomes available.

Despite the achievements presented by the aforementioned studies, several challenges remain in creating robust spoken question answering and information retrieval systems. Works such as CLEAR [9], DAQA [10] and Clotho-AQA [11] require information retrieval from audio segments but do not include open-domain spoken question answering. Much of the work done in spoken question answering is evaluated on datasets such as the Spoken SQuAD dataset [12] or Spoken CoQA dataset [7]. These datasets often only contain spoken questions and contexts that were either generated using text-to-speech or read from a script created from an existing text question answering dataset. This means that further work may be necessary to create spoken language understanding systems that are robust to the disfluencies and lack of proper logical organization often found in spontaneous speech [13]. [14] introduces ODSQA, a spoken question answering dataset of short Chinese utterances. Many of these works format the problem of spoken question answering as finding an answer from a short context (e.g. a one-minute audio recording), or primarily contain non-English speech. Many contexts (e.g. a lecture, an instructional video, or a meeting recording) may be significantly longer, and it is non-trivial to scale a model trained for short contexts to infer answers from a longer context. In addition, further work is needed to ensure that these systems are robust to differences in dialect, accent, speaking style, and regional diction or other out of vocabulary words. This may be especially true for cascade systems employing pre-trained models that were trained on only one dialect or speaking style.

In this work, we aim to advance methods for spoken question answering from long contexts on spontaneous speech. We first introduce the CORAAL Question Answering (CORAAL QA) dataset, which is composed of hand-labeled question-answer-span pairs about information present in long audio files (typically 30min-1hr) from the Corpus of Regional African American Language [15]. Next, we train a model to rank the relevance of segments of ASR transcripts of a long audio file to an input query and return the most likely span to contain a corresponding answer. Finally, we leverage large language models to generate new questions for data augmentation and further process the returned outputs to improve performance.

2. METHODS

We format the question answering from long audio files as the following information retrieval task: an audio file, D, is composed of several short segments, $D = \{s_1, s_2, s_3, ..., s_n\}$. The user inputs a query, q, whose answer, a, can be found in one or more consecutive time segments, i.e. $a = \{s_i, ..., s_{i+k}\}$. Given the input query, q and audio file, D, which may be up to an hour or more in length, we then seek to return the set of answer segment, a. We accomplish this by assigning a score to each segment in D based on its likelihood of answering q and return the segments with the highest scores. For simplicity, we do not consider queries that can not be answered by any segment in the audio file.

2.1. Dataset

To assess performance in this task, we introduce the CORAAL QA dataset [1]. This dataset consists of hand-labeled question-answer pairs created from speech contained in the LES, ROC, DCB, PRV, and VLD splits of the Corpus of Regional African American Language. These splits contain 143 audio files total, of which 34 are under 15min in length, 39 are between 15min and 45min in length, and 70 are greater than 45min in length. The recordings in this dataset consist of informational sociolinguistic interviews between an interviewer and a speaker of a regional variant of African American English. In each interview, the participant is asked questions about their culture, experiences, and opinions. There are 65 different interviewees in total, ranging in age from 12 years old to over 80 years old. From each interview recording, the authors of this paper created a set of questions using the following criteria: 1) The question can be factually and objectively answered by information contained in a continuous time span of the audio file that is 45sec or less in length, 2) The answer to the question is given only once in the audio file, and 3) the answer to the question is not common knowledge and must be answered through extraction from the given audio file. The question answer pairs are given in the format: "query: answer_start_span, answer_end_span" where the answer starting and ending span are given in seconds (e.g. "Who is the speaker's favorite basketball player? : 831.25, 842.76" where the numbers after the colon indicate the start and stop time in the audio file where the speaker gives the answer to the question).

2.2. Model

An overview of the framework is given in Figure 1. The input audio file is first divided into short segments with an overlap between them. To identify the ideal segment size and overlap, we validate over several choices (shown in Table 3) and arrive at an ideal segment size of 60 sec and overlap size of 20 sec. The input audio segments are then transcribed using the Whisper-Large [16] ASR model. Prior work shows that Whisper achieves state-of-the-art performance for the African American English contained in the CORAAL database (16.2% WER) [16, 17]. From the ASR transcript generated from each audio segment, we then use Sentence-BERT [18] to compute a sentence embedding. BERT-based sentence embeddings have been shown to be useful for separating information by topic relevance in text information retrieval tasks [19], and so we seek to apply those benefits in the spoken domain. Inspired by the popular speaker embedding approach of [20], we train a Probabilistic Linear Discriminant Analysis (PLDA) [21] model to score the relevance between the 384-dim BERT sentence embeddings from an input query and the BERT sentence embeddings from the segment-level ASR transcripts. During training, embeddings of text questions from the training set and ASR transcripts from the corresponding answer segments are labeled as coming from the same distribution. During inference, we then use embeddings of target questions as the enrollment set and embeddings of segment-level ASR transcripts as the test set. We then evaluate the system performance by the equal error rate (EER) as well as the precision, recall, and F1-score in correctly retrieving the relevant audio segments. For calculating precision and recall, the PLDA scores for each segment are converted to binary detection decisions by thresholding at the equal error rate. We then compare these scores to the ground truth segment-level labels.

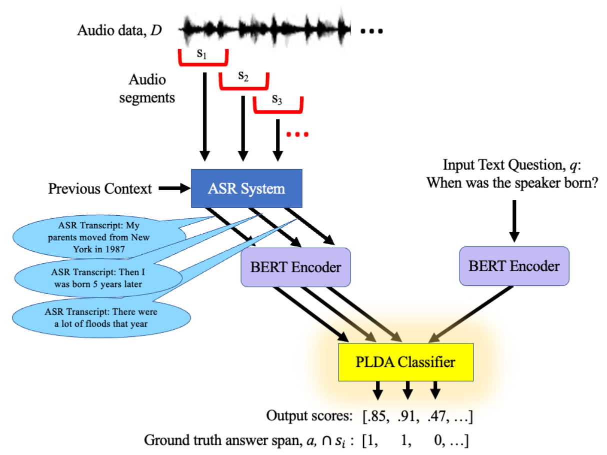
Audio data, D
...
s_1
s_2
s_3
Audio segments
...
Previous Context
ASR System
Input Text Question, q: When was the speaker born?
ASR Transcript: My parents moved from New York in 1987
ASR Transcript: Then I was born 5 years later
ASR Transcript: There were a lot of floods that year
BERT Encoder
BERT Encoder
PLDA Classifier
Output scores: [.85, .91, .47, …]
Ground truth answer span, a, $\cap s_i$: [1, 1, 0, …]

Fig. 1. Overview of the proposed system. The long audio file, D is segmented into one minute segments, s_i. Each segment is then transcribed with ASR, where the ASR system is prompted with previous context. Then both the ASR transcript from each segment and the text of an input query, q, are encoded with Sentence-BERT and scored for the likelihood that s_i answers q by the PLDA classifier. Last, the ground truth scores are used to evaluate performance.

2.3. Experiments

We use the VLD split of CORAAL as testing data and the other splits of CORAAL as training and validation splits. This creates an approximately 80%, 10%, 10% split. Splitting data in this way ensures that no speaker appears in more than one split, and that the regional diction and dialect from the test set has not been previously seen by the model. We first establish the performance of the baseline model with this test-train split in Table 1, validating over different input audio segment lengths. Then, we experiment with two methods designed to improve the model performance: Data augmentation and Prompt Engineering. Inference for all models utilized for these tasks is conducted on a single Nvidia A6000 GPU.

2.3.1. Data Augmentation with Question Generation

In order to improve model performance, we investigate using large language models to generate more diverse training data. In addition to the handwritten questions of the training set, we also use question generation with DeBERTa [22], ChatGPT (gpt-3.5-turbo) [23], and Llama 2-7b [24] to generate additional training questions. Each language model is given the Whisper ASR transcript from each segment

[1] data available at https://github.com/christinachance/CORAAL-QA/

of each audio file in the training set. Then, using each segment-level transcript as context, the language models are prompted to generate a question with an extractive answer. This question and corresponding time frame from the audio are then used as additional training data. In order to evaluate the quality of the generated questions, we generate questions from the same context as the hand-written questions and compare them with the following metrics: **Semantic Similarity**: the cosine distance between the BERT sentence embeddings of the hand-written question and the generated question. **Percent Words Shared**: The number of words that the generated and hand-written questions have in common after lemmatization and removal of stop words divided by the number of words in the hand-written question. **BLEU Score**: As the BLEU score is a commonly accepted metric for the quality of a machine-generated sentence with respect to a human-written sentence, we report the BLEU of the generated question with respect to the hand-written question. **Percent entities included**: We report the number of named entities that appear in the generated question divided by the number of entities that appeared in the context from which the question was generated. We perform this both using the ground truth transcript and ASR transcript as context. This metric gives a measure of the language model's ability to ask questions about specific names, dates, locations, and other named entities as well as the ASR system's ability to correctly transcribe these entities before they are passed to the language model as context. **Answer Precision, Recall, and F1 score**: We first ask a RoBERTa model optimized on the SQuAD dataset [25] to extractively answer both the hand-written and the generated question from the ASR transcript. We then score the precision, recall, and F1-score of the retrieved answer of the generated question with respect to that of the hand-written question using the SQuaD evaluation script [26]. This gives a measure of similarity between the answers to the generated questions and the answers to the hand-written question. **Distractor Accuracy**: For each question generated by each language model, we utilize the MQAG framework [27] to generate a correct answer to the question as well as three incorrect distractor answers. We then ask ChatGPT to answer the four-choice multiple response question from the created answers and report the accuracy. We perform this with distractors both generated from the ground truth transcript and from the ASR transcript. **Answerable score**: We use SelfCheckGPT [28] to derive a score for how answerable a given generated question is given the context. This will ideally return a low score if the generated question contains several errors or does not correspond to the given input context. We perform this using both the ground truth and ASR transcripts as input context. These metrics are shown in Table 2. The performance of the data augmentation experiments is shown in Table 3.

For question generation with Llama 2 and ChatGPT, we elect to feed the models with the following prompt:

You are a Teacher. Your task is to setup a question for an upcoming quiz. The question should be simple in nature. Restrict the question to the context information provided
TRANSCRIPT FROM AUDIO SEGMENT

In total, we generated 7347 questions each using ChatGPT and DeBERTa, and 7230 questions from Llama 2, resulting in a combined total of 21924 augmented questions. The discrepancy in the number of questions generated from Llama 2 arises due to safeguards placed in the model that lead to a refusal to generate questions from certain segments covering sensitive topics.

For predicting the right answer to a question from a set of distractors, we feed ChatGPT with the following prompt:

Chunk Size \(Overlap)	Precision ↑	Recall ↑	Macro F1 ↑	EER ↓
15s \(5s)	0.666	0.567	0.59	0.244
30s \(10s)	0.748	0.622	0.654	0.203
60s \(20s)	0.712	0.631	0.655	0.181

Table 1. Effect of the size of the Audio Segments for predictions from the PLDA model. Precision, Recall and Macro F1 statistics are calculated from predicted scores from the system. EER refers to the Equal Error Rate of the trained system

You are a Student. Your task is to select the correct answer in a test. You are provided with some context information, a question and some multiple choice options. Answer the question only with the context information provided.Return only the correct option.
TRANSCRIPT FROM AUDIO SEGMENT
Question is below
GENERATED QUESTION
Options are below
A: **OPTION 1** *B:* **OPTION 2** *C:* **OPTION 3** *D:* **OPTION 4**

2.3.2. Whisper Prompting

Whisper is powerful in its ability to provide context to the decoder in order to improve transcription, and this has led to significant improvements in word error rate for zero-shot spoken language tasks [29]. In this work, we apply Whisper's prompting to take advantage of the temporal relationship of audio segments upon being input into the classifier model. We try prompting Whisper with the concatenation of ASR transcripts from the last N segments upon transcribing the current segment. This is intended to ensure that segments are transcribed with previous knowledge of the conversation and that important information is preserved and consistently transcribed over time. We experiment with using the last N segments as context in the prompt for $N = 1, 2$, and 3 (shown in Table 4).

3. RESULTS

Table 1 shows the performance of the model with input segments of varying length and overlap. We determine that the question answering model performed best (in terms of both F1 score and EER) with input segments that were 60 seconds long with 20 seconds of overlap between adjacent segments, and so we keep these parameters throughout the rest of this paper. Table 2 gives the evaluation metrics for the questions generated by DeBERTa, ChatGPT, and Llama 2 with respect to the hand-written questions. Table 3 reports the performance of the question answering model using generated questions from each language model as training data. Table 4 reports the performance of the question answering model when using N segments of previous context as the prompt to Whisper in the current step.

4. DISCUSSION

As all the language models used in this work are trained on text data, their metrics in generating data for a spontaneous speech task are expected to be lower than if evaluated on written words. However, ChatGPT seems to be more robust to the differences in spoken speech vs written text style than DeBERTa and Llama 2. We observe from Table 2 that ChatGPT often produces questions most in line with the hand-written questions, having the highest semantic similarity, number of shared words, and BLEU score with the human-generated samples. ChatGPT's questions also had higher distractor

Model	Semantic Similarity $\uparrow$	Percent Words Shared $\uparrow$	BLEU $\uparrow$	GT % Entities Included $\uparrow$	Whisper % Entities Included $\uparrow$	Answer (precision/ recall/ f1) $\uparrow$	GT Distractor Acc $\uparrow$	Whisper Distractor Acc $\uparrow$	GT Answerable Score $\uparrow$	Whisper Answerable Score $\uparrow$
DeBERTa	0.3914	28.28	0.063	20.07	14.52	37.1 /35.4/34.2	72.97	71.86	68.62	63.55
ChatGPT	0.5670	43.17	0.065	43.19	20.49	41.3/ 40.8/ 39.2	73.43	72.97	92.88	92.55
Llama 2	0.4751	38.85	0.054	39.61	28.42	30.0/ 29.9/ 28.4	65.28	64.46	91.76	92.00

Table 2. Metrics for evaluating the quality of generated questions: Cosine distance between BERT embeddings of the generated questions and hand-written questions (Semantic Similarity), Percent Words shared between the generated questions and hand-written questions, BLEU score between the generated and hand-written questions, % of named entities from the ground truth transcript not included in the generated question (GT % entities included), % of named entities from ASR transcript included in the generated question (Whisper % entities included)), Precision, Recall, and F1-score between the retrieved answer to the hand-written question and the generated question (Answer precision, recall, and F1-score), language model accuracy in correctly answering the question from a multiple choice set with distractors generated from the GT transcript (GT Distractor Acc) and generated from the ASR transcript (Whisper Distractor Acc), and the Answerable score given by SelfCheckGPT for the generated question with either the GT transcript or the ASR transcript given as context (GT Answerable Score and Whisper Answerable score, respectively).

Model	Precision $\uparrow$	Recall $\uparrow$	Macro F1 $\uparrow$	EER $\downarrow$
DeBERTa	0.732	0.64	0.667	0.183
ChatGPT	0.76	0.66	0.688	0.175
Llama 2	0.748	0.654	0.681	0.164
All	0.765	0.668	0.697	0.166

Table 3. Performance of PLDA systems trained with questions generated by different systems. Questions were generated by the respective systems from the non-prompted Whisper generated ASR transcripts. All refers to a PLDA model trained by combining the questions generated from all the individual models

#Seg	Precision $\uparrow$	Recall $\uparrow$	Macro F1 $\uparrow$	EER $\downarrow$
$N = 1$	0.728	0.641	0.680	0.175
$N = 2$	0.759	0.658	0.688	0.174
$N = 3$	0.761	0.663	0.689	0.175

Table 4. Performance of the question answering model when the ASR transcripts from the previous N segments are used in the Whisper prompt as previous context on transcribing the current audio segment.

accuracy, though we acknowledge that there may be bias in evaluating its performance, as we predict distractor scores through a secondary prompt to ChatGPT itself. The questions generated by DeBERTa give a significantly lower answerable score than those from either ChatGPT or Llama 2. However, when asked to answer both its own generated question and a hand-written question derived from the same audio segment, DeBERTa gives higher answer precision, recall, and F1-score than Llama 2. However, the number of named entities that Llama 2 included in its questions generated from the both the ground truth transcript and the ASR transcript was significantly higher than that from DeBERTa and not significantly different from that from ChatGPT.

Although using generated questions from ChatGPT in training improved the performance (in terms of precision, recall, and F1 score) over the baseline without data augmentation, we see that additional training data generated by DeBERTa and Llama 2 was also beneficial. This may imply that using artificially generated questions in the question-answering system is beneficial regardless of the generative model used, although some models may produce more human-like or diverse sets of training samples. It also appears that the semantic similarity and Whisper answerable score metrics of the generated questions correlate relatively well to the precision, recall, and F1 score of the question answering model trained on that synthetic data. These metrics may be useful for data mining or automatic quality analysis of generated data in the future. Next, the combination of questions generated by DeBERTa, ChatGPT, and Llama 2 together in data augmentation gives a larger benefit than the use of generated questions from any one model. This result may mean that having a combination of questions from different models, (i.e. a more diverse augmented training set) is more important than the quality of questions generated by any one model.

The model appears to benefit from the additional context in the prompt. The system shows the highest F1 score when using $N = 3$ segments of previous context, implying that the model performs better when given more context. However, the increase in performance from $N = 2$ to $N = 3$ is marginal, and the benefits gained would likely be outweighed by the additional memory cost if significantly more audio segments were used.

5. CONCLUSIONS

This paper proposes a promising framework for information retrieval from long audio files for question answering. The framework, which uses PLDA to score the relevance of BERT-based embeddings of ASR transcript segments to an input query, is further enhanced by two methods 1) A data augmentation method that leverages large language models to generate synthetic training data and 2) effective prompt engineering in the ASR system to allow utilization of temporal information and consistent transcription across segments. Using the CORAAL QA database for spoken question answering from spontaneous speech that we introduce here, we are able to show good performance in the spoken language task of retrieving information from long audio files as well as demonstrate the effectiveness of large language models and prompt engineering in improving the model's accuracy. Future steps include combining the data augmentation and prompt generations proposed here to further improve the system output, and additional studies on how large language models can be adapted or prompted to produce more spontaneous speech-like utterances for enhanced data augmentation of spoken language systems. Future work also includes showing how well the model trained on the CORAAL QA dataset generalizes to spoken speech in other domains (different accents, speaker styles, etc.) [2].

[2] This work is supported in part by the NSF

6. REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 2019.

[2] Alec Radford, Jeffrey Wu, et al., “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, pp. 9, 2019.

[3] Nuo Chen, Chenyu You, and Yuexian Zou, “Self-Supervised Dialogue Learning for Spoken Conversational Question Answering,” *Proceedings Interspeech 2021*, pp. 231–235, 2021.

[4] Suwon Shon, Siddhant Arora, et al., “SLUE phase-2: A benchmark suite of diverse spoken language understanding tasks,” *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023.

[5] Yung-Sung Chuang, Chi-Liang Liu, et al., “SpeechBERT: An Audio-and-Text Jointly Learned Language Model for End-to-End Spoken Question Answering,” *Proceedings Interspeech 2020*, pp. 4168–4172, 2020.

[6] Xuxin Cheng, Zhihong Zhu, et al., “GhostT5: Generate More Features with Cheap Operations to Improve Textless Spoken Question Answering,” *Proceedings Interspeech 2023*, pp. 1134–1138, 2023.

[7] Chenyu You, Nuo Chen, et al., “End-to-end spoken conversational question answering: Task, dataset and model,” *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022.

[8] Guan-Ting Lin, Yung-Sung Chuang, et al., “DUAL: Discrete Spoken Unit Adaptive Learning for Textless Spoken Question Answering,” *Proceedings Interspeech 2022*, pp. 5165–5169, 2022.

[9] Jerome Abdelnour, Giampiero Salvi, and Jean Rouat, “Clear: A dataset for compositional language and elementary acoustic reasoning,” *arXiv*, vol. abs/1811.10561, 2018.

[10] Haytham M. Fayek and Justin Johnson, “Temporal reasoning via audio question answering,” *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2283–2294, 2020.

[11] Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen, “Clotho-aqa: A crowdsourced dataset for audio question answering,” *30th European Signal Processing Conference, EUSIPCO 2022*, pp. 1140–1144, 2022.

[12] Chia-Hsuan Lee, Szu-Lin Wu, et al., “Spoken SQuAD: A Study of Mitigating the Impact of Speech Recognition Errors on Listening Comprehension,” *Proceedings Interspeech 2018*, pp. 3459–3463, 2018.

[13] S. Furui, T. Kikuchi, et al., “Speech-to-text and speech-to-speech summarization of spontaneous speech,” *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 4, pp. 401–408, 2004.

[14] Chia-Hsuan Lee, Shang-Ming Wang, Huan-Cheng Chang, and Hung-yi Lee, “ODSQA: open-domain spoken question answering dataset,” *2018 IEEE Spoken Language Technology Workshop, SLT*, pp. 949–956, 2018.

[15] T. Kendall and C. Farrington, “The Corpus of Regional African American Language. Version 2021.07.,” 2021.

[16] Alec Radford, Jong Wook Kim, et al., “Robust speech recognition via large-scale weak supervision,” *arXiv*, vol. abs/2212.04356, 2022.

[17] Alexander Johnson, Hariram Veeramani, et al., “An equitable framework for automatically assessing children’s oral narrative language abilities,” *Proceedings Interspeech 2023*, 2023.

[18] Nils Reimers and Iryna Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[19] Min Pan, Junmei Wang, et al., “A probabilistic framework for integrating sentence-level semantics via bert into pseudo-relevance feedback,” *Information Processing & Management*, vol. 59, no. 1, pp. 102–134, 2022.

[20] David Snyder, Daniel Garcia-Romero, et al., “X-vectors: Robust dnn embeddings for speaker recognition,” *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[21] Sergey Ioffe, “Probabilistic linear discriminant analysis,” *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Proceedings, Part IV 9*, pp. 531–542, 2006.

[22] Pengcheng He, Xiaodong Liu, et al., “Deberta: Decoding-enhanced bert with disentangled attention,” *International Conference on Learning Representations*, 2021.

[23] OpenAI, “Gpt-4 technical report,” *arXiv*, vol. abs/2303.08774, 2023.

[24] Hugo Touvron et al., “Llama 2: Open foundation and fine-tuned chat models,” *arXiv*, vol. abs/2307.09288, 2023.

[25] Xiang Pan, Alex Sheng, et al., “Task transfer and domain adaptation for zero-shot question answering,” *arXiv*, vol. abs/2206.06705, 2022.

[26] Pranav Rajpurkar, Jian Zhang, et al., “SQuAD: 100,000+ questions for machine comprehension of text,” *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

[27] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales, “MQAG: multiple-choice question answering and generation for assessing information consistency in summarization,” *arXiv*, vol. abs/2301.12307, 2023.

[28] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales, “Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models,” *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pp. 9004–9017, 2023.

[29] Puyuan Peng, Brian Yan, et al., “Prompting the Hidden Talent of Web-Scale Speech Models for Zero-Shot Task Generalization,” *Proceedings Interspeech 2023*, pp. 396–400, 2023.