Data-Centric Improvements for Enhancing Multi-Modal Understanding in Spoken Conversation Modeling

Maximillian Chen[*]
Columbia University and Google
maxchen@cs.columbia.edu

Ruoxi Sun
Google
ruoxis@google.com

Sercan Ö. Arık
Google
soarik@google.com

Abstract

Conversational assistants are increasingly popular across diverse real-world applications, highlighting the need for advanced multimodal speech modeling. Speech, as a natural mode of communication, encodes rich user-specific characteristics such as speaking rate and pitch, making it critical for effective interaction. Our work introduces DAMSEL, a data-centric customization approach for efficiently enhancing multimodal understanding in conversational speech modeling. Central to our contributions is a novel multi-task learning paradigm that involves designing auxiliary tasks to utilize a small amount of speech data. Our approach achieves state-of-the-art performance on the Spoken-SQuAD benchmark, using only 10% of the training data with open-weight models, establishing a robust and efficient framework for audio-centric conversational modeling. We also introduce ASK-QA, the first dataset for multi-turn spoken dialogue with ambiguous user requests and dynamic evaluation inputs.

1 Introduction

Real-world adoption of intelligent multimodal conversational agents has progressed quickly in recent years due to the impressive capabilities of Large Language Models (LLMs). However, despite numerous applications, including smart home systems, contact centers, customer support/service, personalized education etc. (Hemphill et al., 1990; Khatri et al., 2018b; Li et al., 2017; Von Ahn, 2013; Fatima et al., 2024; Li et al., 2023; Zheng et al., 2024), there has not been the same rapid progress in adapting Multimodal LLMs (MLLMs) to spoken contexts due to several fundamental challenges.

Speech data constitute high-dimensional signals that are difficult to model even for frontier models (e.g., Whisper-based models are limited to 30 seconds; Chu et al. (2024); Radford et al. (2023)) and Gemini 1.5 represents 1 second of audio using 25 tokens[1]). These are temporal signals which include acoustic phenomena (e.g. background noise Mehrish et al. (2023)) and important paralinguistic information such as speaking rates or pitch (Hirschberg, 1993; Bhattacharya et al., 2023). Performance on speech understanding tasks is thus greatly affected by the ability to robustly comprehend the semantic contents of the input speech (Li et al., 2017), as illustrated in Fig. 1. This is further complicated by the long-standing issue of models overfitting to individual speakers (Jung et al., 2018; Wang et al., 2020). These can be viewed as a shortcoming of insufficient training data coverage (Yang et al., 2024b). However, large-scale speech data collection is notoriously difficult due to privacy concerns (Nautsch et al., 2019; Qian et al., 2018).

Despite the difficulty of large-scale collection, task-specific data is increasingly the most effective approach to guarantee use-case customization for state-of-the-art MLLMs like Gemini or GPT (Gem-

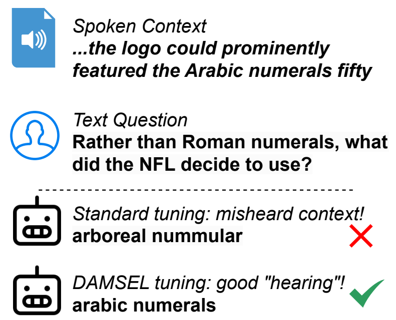

Figure 1: Automatic speech recognition is a necessary *implicit* skill for MLLM in end-to-end spoken question answering. We propose DAMSEL, a data-centric multi-task learning approach which explicitly teaches these skills, as exemplified by this QA pair from Spoken-SQuAD.

[*]Work done during an internship at Google.

[1]https://ai.google.dev/gemini-api/docs/audio?lang=python

Findings of the Association for Computational Linguistics: ACL 2025, pages 1366–1387
July 27 - August 1, 2025

ini Team et al., 2023; Brown et al., 2020). These models are closed-source, but offer commercial tuning APIs, which typically do not permit modifications to the model or learning objective. Even with smaller open-weight models, it can still be computationally intractable to iterate on architectures and train from scratch due to the expensive compute resource demands (Groeneveld et al., 2024). These motivate the design of efficient data-centric methods (Seedat et al., 2022) which maximize models' ability to overcome the aforementioned challenges of long speech understanding reliably.

In this work, we take a data-centric perspective towards addressing the varied challenges of adapting multimodal LLMs for speech. Our contributions can be summarized as follows:

- We bring a multi-task learning paradigm to improve speech understanding implemented via a simple but effective data-centric approach called **DAMSEL** (**DA**ta-centric **M**ulti-task **S**p**E**ech **L**earning). Rather than using additional datasets, we instead design auxiliary tasks to maximize cross-modal learning from a fixed set of recorded speech.
- We propose **A**mbiguous **S**po**k**en Conversational **Q**uestion **A**nswering (ASK-QA), a novel dataset which combines the challenges of multimodal speech modeling and mixed-initiative interaction. ASK-QA features contextually ambiguous questions along with long multi-turn speech and diverse accents, speaking rates, and pitches.
- We validate the proposed data-centric approach on three spoken question answering (SQA) corpora: ASK-QA, Spoken-SQuAD, and SD-QA, representing various combinations for whether input questions and knowledge context are represented as text or speech. Our approach applied even to open-weight models is able to outperform the existing state-of-the-art on Spoken-SQuAD using only 10% of the available training data.

2 DAMSEL: Data-centric, Multi-task Speech Learning

We consider the setting of customizing an MLLM for use in request-based end-to-end speech modeling, similarly to Shih et al. (2024). An MLLM is provided as input an audio recording and textual context. The backbone of many MLLM architectures is a textual decoder-only LLM (Liu et al., 2024), so the textual context usually contains an instruction. These settings involve reasoning about some contextual knowledge and conversation history. The model aims to provide a correct answer to a target question (i.e. the last conversation turn). Different applications may involve spoken conversations about written information (e.g. document-grounded QA), or written conversations about spoken information (e.g. meeting summarization).

Tuning MLLMs with cross-entropy loss is advantageous as it can be used to unify diverse tasks as a single text-to-text objective (Raffel et al., 2020). Many recent studies find that multi-task learning (Caruana, 1997) using *additional datasets* greatly improves downstream task performance (Aghajanyan et al., 2021; Padmakumar et al., 2022; Chen and Yu, 2023). Recent MLLM efforts such as Qwen2-Audio or SALMONN rely on similar intuitions by training their models on many large-scale audio datasets prior to downstream use (Chu et al., 2024; Tang et al., 2024). *Here, we design auxiliary tasks within the same dataset to maximize cross-modal learning gains from a fixed set of audio recordings for a target dataset, rather than collecting any additional recordings for generic improvements.* DAMSEL involves breaking down our problem into three intermediate goals: 1) correctly representing the spoken context, 2) learning to reason across all input modalities, and 3) coherently producing the correct answer.

1) Listening Comprehension is an auxiliary task to help the SLM "hear" the spoken context. It has been consistently reported in traditional cascade-style systems that SQA performance is greatly affected by automatic speech recognition (ASR) errors (Li et al., 2018), and thus we design a task to specifically address this point. The objective is for the MLLM to predict a ground-truth (or pseudo-labeled) audio transcription, given a recording and a task instruction as input.

2) Cross-Modal Commonsense Reasoning is an auxiliary task designed to unify the contents of the spoken and textual inputs. We reframe dialogue response selection (Henderson et al., 2019) as a multiple-choice reasoning task (Talmor et al., 2019). The answer options consist of the correct answer (e.g. "It was recovered a few months later") and commonsense negative answer choices sampled from other training QA pairs (e.g. "Do you mean the popular generic name?"), as shown in

Table A8. The objective is to solve this multiple-choice reasoning task by selecting the correct answer given the recording, conversation context, knowledge, answer options, and a task instruction. This auxiliary task is intentionally simple (the incorrect choices are each easy negatives) as the goal is to supplement MLLMs' cross-modal reasoning abilities which may not be explicitly learned from the ingested data during pre-training. This approach is also inspired by work in contrastive learning which seeks to highlight the benefits of correct selections (e.g., the data generation phase in Chen et al. (2025)).

3) Response Generation is the primary objective of providing a correct answer. The inputs are what is expected to be provided to an MLLM at inference time for SQA: the recording, conversation context, information context, and a task-specific instruction. An example is shown in Table A11.

These tasks can be fully implemented as modifications to tuning data mixtures. As shown in Sec. 3, this simple modification is observed to be highly effective in improving an MLLMs' ability to complete downstream tasks, particularly in the limited data regime. We provide implementation details for our approach in Appendix D.1.

	SD-QA	S-SQuAD	ASK-QA (Ours)
Avg. Audio Len.	4.8s	59s	1m 41s
Speakers/Audio	1	1	3
Knowledge	Text	Speech	Speech
Conversation	Speech	Text	Speech
Unique Voices	248	1	64
Avg. Turns	2	2	5.1
Answer Type	Span	Span	Free-form
Ambiguous	✗	✗	✓
Dynamic Eval	✗	✗	✓
Disfluencies	✗	✗	✓

Table 1: **Comparison of ASK-QA against existing popular SQA training datasets used for experimentation here**. ASK-QA features ambiguous requests and long audio context. See Appendix 3.3 for SD-QA.

3 Experiments

We evaluate our approach on SQA datasets with different combinations of context modalities (see Table 1). We prompt and fine-tune two closed-source models, Gemini Pro and Gemini Flash, on the Vertex AI platform[2]. We also use Speech-Qwen, which we built by pre-training an 17.8M parameter projection layer between a frozen audio encoder (WavLM-Large) and a frozen LLM

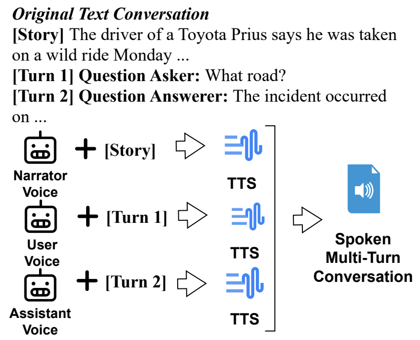

Figure 2: **Simplified summary of the pipeline for constructing ASK-QA.** For each text conversation in Abg-CoQA, we construct three speaker profiles with randomly sampled voices, speaking rates, and pitches. We use TTS to synthesize the story context as a spoken narration, then each individual dialogue turn. The resulting audio files are joined as a single recording.

decoder (Qwen 2.5 7B-Instruct). See Appendix C for details on Speech-Qwen. Our main findings for ASK-QA and Spoken-SQuAD are reported in this section. Full results and extended details for each experimental setting are reported in Appendix E.

3.1 ASK-QA: Spoken Knowledge and Multi-Turn Spoken Dialogue

We develop a novel corpus for speech-based mixed-initiative conversation: **A**mbiguous **Sp**oken **C**onversational **Q**uestion **A**nswering (ASK-QA). The contextual inputs for ASK-QA are *fully spoken*.

Dataset Construction: We construct ASK-QA starting from Abg-CoQA (Guo et al., 2021), a span-based textual conversational QA task. Given a story as context, each conversation consists of dialogue turns where a user asks questions and an assistant is supposed to provide the correct answer or ask a clarifying question if the user's request is ambiguous. Our data construction pipeline is summarized in Figure 2. The data generation process is detailed in Appendix B. In total, ASK-QA contains 221.8h of speech. The training, validation, and test sets contain 5,985, 500, and 1,345 conversations. The average audio input length of ASK-QA is 1 minute and 41 seconds (see Table 1), which surpasses the input length of many MLLMs such as Qwen2-Audio (Chu et al., 2024).

Evaluation: Following recent work (Chen et al., 2025; Risch et al., 2021), we apply embedding-based semantic similarity (Reimers, 2019) to allow for flexible free-form QA evaluation. We apply this metric to a standard single-turn setting as well as a novel multi-turn setting which combines TTS

[2]https://cloud.google.com/vertex-ai/generative-ai/docs/models/tune_gemini/audio_tune

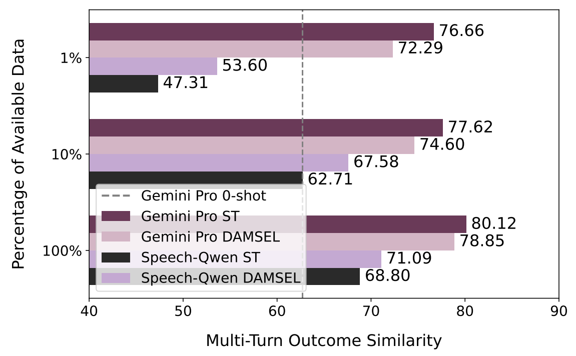

Figure 3: DAMSEL, our multi-task tuning approach, improves upon Single-task (ST) tuning with Gemini and Speech-Qwen on ASK-QA's multi-turn evaluation.

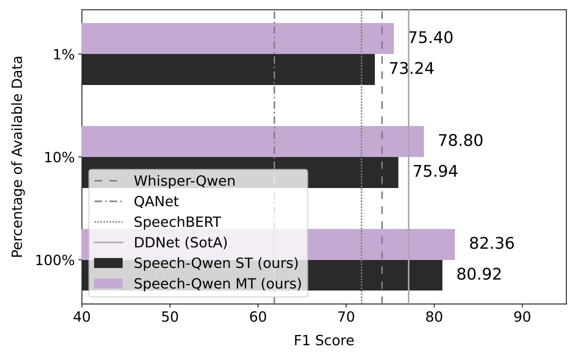

Figure 4: DAMSEL applied to Speech-Qwen outperforms the state-of-the-art approach on Spoken-SQuAD using only 10% of the available data.

with the dynamic input evaluation for Abg-CoQA in Chen et al. (2025). See details in Appendix B.2.

Findings: We benchmark end-to-end performance on ASK-QA in Table A6 using DAMSEL (as described in Section 2) and baseline single-task tuning (which represents standard end-to-end speech-to-text modeling (Shih et al., 2024)). The listening comprehension sub-task separately models the story and conversation transcripts, inspired by speaker diarization (Anguera et al., 2012; Gu et al., 2021; Yu et al., 2022). With Speech-Qwen, we see as much as 13.3% relative improvement over standard fine-tuning depending on the amount of available data on trajectory-level similarity in Figure 3. Surprisingly, with Gemini Pro, we also see relative improvements of 5.7% with 1% of the available training data and 1.6% when using full data, despite frontier MLLMs already having large-scale multi-modal pre-training and the full ASK-QA corpus containing large-scale, in-distribution data (over 200 hours). This finding is significant because it *specifically indicates that the MLLM is better learning to model the available speech data.* It is well-documented that 1) high benchmark scores achieved by frontier LLMs on older corpora may be confounded by data contamination (Roberts et al., 2023; Qian et al., 2024), and 2) several studies demonstrate the efficacy of direct fine-tuning given abundant data (Sharma et al., 2024; Yu et al., 2024). Since ASK-QA is newly synthesized, Gemini cannot have been trained on this exact version of the data. This accurately highlights the difference between direct single-task tuning and multi-task tuning with our proposed auxiliary tasks. The improvements with full data indicate the applicability of the approach for scaled data.

3.2 Spoken-SQuAD: Spoken Knowledge and Textual Questions

Spoken-SQuAD (Li et al., 2018) is a speech version of SQuAD (Rajpurkar et al., 2016). Rather than span-based classification, we solve the task using our end-to-end generative approach. Each instance has a textual question and spoken knowledge.

Findings: In Figure 4, we benchmark DAMSEL against single-task tuning via Speech-Qwen. Our performance is evaluated in terms of exact match and F1 score using the SQuAD evaluator. Our approach, applied to an open-weight model like Speech-Qwen, outperforms the existing state-of-the-art model proposed in You et al. (2022) using just 10% of the available training data, indicating that it is *highly efficient and effective for cross-modal learning*. Our expanded results which include an additional MLLM are shown in Table A7. We see similar trends (up to 52.8% improvement given limited data) on the multi-lingual SD-QA corpus in Section 3.3.

3.3 SD-QA: Textual Knowledge and Spoken Questions

We examine the setting where the single-turn QA context is provided in the recorded speech, and the knowledge necessary to answer the question correctly is provided in the text.

3.3.1 Dataset

SD-QA (Faisal et al., 2021) is a large single-turn SQA benchmark with diverse data – spanning 5 languages (Arabic, Bengali, English, Kiswahili, and Korean) and 24 regional dialects. SD-QA is also proposed as a span-based QA task, but we apply our end-to-end generative approach as in

Base Model	Objective	Data	EM ↑	F1 ↑
Gemini Pro	–	0%	42.44	64.18
Gemini Pro	Single	1%	46.33	67.74
Gemini Pro	DAMSEL	1%	**55.24**	**70.73**
Gemini Pro	Single	10%	62.79	77.80
Gemini Pro	DAMSEL	10%	**63.04**	**79.02**
Gemini Pro	Single	100%	63.10	78.15
Gemini Pro	DAMSEL	100%	**64.17**	**79.06**
Speech-Qwen	Single	1%	13.44	25.28
Speech-Qwen	DAMSEL	1%	**24.70**	**38.63**
Speech-Qwen	Single	10%	29.70	43.84
Speech-Qwen	DAMSEL	10%	**39.92**	**54.35**
Speech-Qwen	Single	100%	46.83	61.76
Speech-Qwen	DAMSEL	100%	**49.54**	**64.94**

Table 2: **Experimental results comparing single-task SFT and our proposed multi-task approach on SD-QA's test set**.

Section 3. We tune our models on up to 9,008 of the 10,0008 samples made available for training, withholding the remaining samples for validation. We evaluate our approach on the 12,975 evaluation samples.

Findings: We evaluate performance on SD-QA in terms of exact match and token-based F1. Consistent with our findings in Section 3, we see that our multi-task approach is able to outperform single-task tuning in all evaluation settings. This is inclusive of experiments with Gemini Pro as the base MLLM for tuning. We see a large 16.13% relative improvement (46.33 to 55.24) for exact match in the limited data setting with Gemini.

We observe that Gemini Pro is already a strong base MLLM, achieving competitive zero-shot performance on this corpus. This is likely due to it having a strong initialization on multilingual ASR. We see that our Speech-Qwen model is able to outperform zero-shot Gemini using our multi-task approach with full data. We also observe up to a 52.8% relative improvement over single-task tuning with Speech-Qwen in the 1% data regime. This is consistent with findings from Chen and Yu (2023) in which pre-finetuning yields strong improvements in the extremely limited data regime. Overall, the particularly large performance improvements on this particular corpus may be an indication that the base models have not been trained on as much multi-lingual data.

3.4 Ablation Studies

In Table 3, we systematically remove each individual task: Dialogue Listening Comprehension (DLC), Story Listening Comprehension (SLC), and Response Selection (RS). The removal of each auxiliary task results in performance degradation relative to full multi-task tuning, indicating their importance towards improved cross-modal understanding. We observe that removing SLC results in the most performance degradation, which follows the intuition in Figure 1 and Li et al. (2018).

Approach	Data	Single-Turn Sim. ↑	Multi-Turn Sim. ↑
DAMSEL w/o DLC	1%	53.77	53.10
DAMSEL w/o SLC	1%	52.32	51.89
DAMSEL w/o RS	1%	53.53	52.67
DAMSEL	1%	54.54	53.60
DAMSEL w/o DLC	10%	65.09	64.19
DAMSEL w/o SLC	10%	64.75	64.24
DAMSEL w/o RS	10%	66.89	66.01
DAMSEL	10%	68.27	67.58

Table 3: Systematic ablations of each individual task type on ASK-QA in the limited data setting using Speech-Qwen.

4 Conclusion

We propose a data-centric multi-task learning approach which helps improve speech data utilization for MLLM tuning. Tuning on various corpora with Gemini and open-weight MLLMs, we observe consistent performance improvements regardless of model scale and tuning access, surpassing state-of-the-art performance on Spoken-SQuAD with open-weight MLLMs. Future work may build upon our insights by designing new auxiliary tasks, incorporating more expressive TTS approaches (e.g., emotion modeling), or examining action optimization strategies for ASK-QA. Our dataset and synthesis process can also be contributed to post-training data mixtures to improve construction of MLLMs' for improved long-context speech modeling abilities.

Acknowledgements

We thank Jinsung Yoon and Ta-Chung Chi for their helpful feedback on our work.

Limitations

Transcription supervision: One of the crucial auxiliary tasks in our approach is listening comprehension, as demonstrated by the performance degradation in our ablations (Table 3). In our implementation, we use ground-truth transcriptions as the target for this generation task. These transcriptions may not be available — for instance, the ones provided by Spoken-SQuAD and SD-QA were obtained via ASR. Our transcription for ASK-QA is not guaranteed to perfectly match the generated speech either, despite our efforts to filter the data quality (see Appendix B). It is not clear whether the possibility of slight transcription errors improves model robustness to noise or degrades performance, and this warrants further study in future work.

TTS quality: Our data generation approach is bottlenecked by current capabilities of TTS software. While TTS has greatly improved in recent years in terms of WER, we do still witness generation errors and naturalness issues when working with long context (hence the need for filtering). We are also not at the point in which we have perfect controllability in paralinguistic attributes.

Generalization to paralinguistic tasks: We propose a multi-task approach which can be used to greatly improve performance in SQA. In the three corpora here, listening comprehension proves to be crucial as the primary objective is auditory semantic understanding. However, in more nuanced contexts like task guidance (Schlager and Feiner, 2024), it is more important to monitor different paralinguistic aspects of the user such as frustration.

Use in large-scale model post-training: We believe that our overall data generation process can be useful for improving MLLM post-training. However, verifying this claim is beyond the scope of this work due to computational constraints. We see improved performance on our downstream task after supervised fine-tuning of Gemini, which does indicate positive signal that there are correlations between our training and evaluation data.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811.

Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. 2012. Speaker diarization: A review of recent research. *IEEE Transactions on audio, speech, and language processing*, 20(2):356–370.

Siddhant Arora, Hayato Futami, Jee-weon Jung, Yifan Peng, Roshan Sharma, Yosuke Kashiwagi, Emiru Tsunoo, Karen Livescu, and Shinji Watanabe. 2024. Universlu: Universal spoken language understanding for diverse tasks with natural language instructions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2754–2774.

Debasmita Bhattacharya, Jie Chi, Julia Hirschberg, and Peter Bell. 2023. Capturing formality in speech across domains and languages. *Interspeech 2023*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28:41–75.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Andy Rosenbaum, Seokhwan Kim, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2022a. Weakly supervised data augmentation through prompting for dialogue understanding. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*.

Maximillian Chen, Ruoxi Sun, Tomas Pfister, and Sercan Ö Arık. 2025. Learning to clarify: Multi-turn conversations with action-based contrastive self-training. *International Conference on Learning Representations*.

Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, and Zhou Yu. 2023. Controllable mixed-initiative dialogue generation through prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 951–966.

Maximillian Chen and Zhou Yu. 2023. Pre-finetuning for few-shot emotional speech recognition. In *INTERSPEECH*. INTERSPEECH.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022b. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yung-Sung Chuang, Chi-Liang Liu, Hung-yi Lee, and Lin-shan Lee. 2020. Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering. *Interspeech 2020*.

Yang Deng, Wenqiang Lei, Wenxuan Zhang, Wai Lam, and Tat-Seng Chua. 2022. Pacific: Towards proactive conversational question answering over tabular and textual data in finance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6970–6984.

Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 807–818.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. Sd-qa: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315.

Kaleem Fatima, Pinky Singh, Hetal Amipara, and Ganesh Chaudhary. 2024. Accuracy of artificial intelligence-based virtual assistants in responding to frequently asked questions related to orthognathic surgery. *Journal of Oral and Maxillofacial Surgery*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, A. Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Daniel Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hanna Hajishirzi. 2024. Olmo: Accelerating the science of language models. *arXiv preprint*.

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. Mpc-bert: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3682–3692.

Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. Abg-coqa: Clarifying ambiguity in conversational question answering. In *3rd Conference on Automated Knowledge Base Construction*.

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Matthew Henderson, Ivan Vulić, Daniela Gerz, Iñigo Casanueva, Paweł Budzianowski, Sam Coope, Georgios Spithourakis, Tsung-Hsien Wen, Nikola Mrkšić, and Pei-Hao Su. 2019. Training neural response selection for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5392–5404.

Julia Hirschberg. 1993. Pitch accent in context predicting intonational prominence from text. *Artificial Intelligence*, 63(1-2):305–340.

Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166.

HY Huang. 2017. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. *arXiv preprint arXiv:1711.07341*.

Jee-weon Jung et al. 2018. Avoiding speaker overfitting in end-to-end dnns using raw waveform for text-independent speaker verification. In *Interspeech*.

Chandra Khatri, Behnam Hedayatnia, Anu Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, et al. 2018a. Advancing the state of the art in open domain dialog systems through the alexa prize. *arXiv preprint arXiv:1812.10757*.

Chandra Khatri, Anu Venkatesh, Behnam Hedayatnia, Raefer Gabriel, Ashwin Ram, and Rohit Prasad. 2018b. Alexa prize—state of the art in conversational ai. *AI magazine*, 39(3):40–55.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, et al. 2023. Soda: Million-scale dialogue distillation with social commonsense contextualization. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *Forty-first International Conference on Machine Learning*.

Gakuto Kurata, Nobuyasu Itoh, and Masafumi Nishimura. 2011. Acoustic model training with detecting transcription errors in the training data. In *INTERSPEECH*, pages 1689–1692.

Chia-Hsuan Lee, Yun-Nung Chen, and Hung-Yi Lee. 2019. Mitigating the impact of speech recognition errors on spoken question answering by adversarial domain adaptation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7300–7304. IEEE.

Bo Li, Tara N Sainath, Arun Narayanan, Joe Caroselli, Michiel Bacchiani, Ananya Misra, Izhak Shafran, Hasim Sak, Golan Pundak, Kean K Chin, et al. 2017. Acoustic modeling for google home. In *Interspeech*, pages 399–403.

Chen Li, Dimitris Chrysostomou, and Hongji Yang. 2023. A speech-enabled virtual assistant for efficient human–robot interaction in industrial environments. *Journal of Systems and Software*, 205:111818.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *arXiv preprint arXiv:1804.00320*.

Yu Li, Kun Qian, Weiyan Shi, and Zhou Yu. 2020. End-to-end trainable non-collaborative dialog system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8293–8302.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, et al. 2024. An embarrassingly simple approach for llm with strong asr capacity. *arXiv preprint arXiv:2402.08846*.

Ambuj Mehrish, Navonil Majumder, Rishabh Bharadwaj, Rada Mihalcea, and Soujanya Poria. 2023. A review of deep learning techniques for speech processing. *Information Fusion*, 99:101869.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797.

Matthew Mulholland, Melissa Lopez, Keelan Evanini, Anastassia Loukina, and Yao Qian. 2016. A comparison of asr and human errors for transcription of non-native spontaneous speech. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5855–5859. IEEE.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. 2024. Spoken question answering and speech continuation using spectrogram-powered llm. In *12th International Conference on Learning Representations, ICLR 2024*.

Andreas Nautsch, Abelino Jiménez, Amos Treiber, Jascha Kolberg, Catherine Jasserand, Els Kindt, Héctor Delgado, Massimiliano Todisco, Mohamed Amine Hmani, Aymen Mtibaa, et al. 2019. Preserving privacy in speaker and speech characterisation. *Computer Speech & Language*, 58:441–480.

Vishakh Padmakumar, Leonard Lausen, Miguel Ballesteros, Sheng Zha, He He, and George Karypis. 2022. Exploring the role of task transferability in large-scale multi-task learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2542–2550.

Tatiana Passali, Thanassis Mavropoulos, Grigorios Tsoumakas, Georgios Meditskos, and Stefanos Vrochidis. 2022. LARD: Large-scale artificial disfluency generation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2327–2336, Marseille, France. European Language Resources Association.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Gueorgui Pironkov et al. 2016. Speaker-aware long short-term memory multi-task learning for speech recognition. In *EUSIPCO*, pages 1911–1915. IEEE.

Jianwei Qian, Feng Han, Jiahui Hou, Chunhong Zhang, Yu Wang, and Xiang-Yang Li. 2018. Towards privacy-preserving speech data publishing. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 1079–1087. IEEE.

Kun Qian, Shunji Wan, Claudia Tang, Youzhi Wang, Xuanming Zhang, Maximillian Chen, and Zhou Yu. 2024. VarBench: Robust language model benchmarking through dynamic variable perturbation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16131–16161, Miami, Florida, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023.

Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. 2023. To the cutoff... and beyond? a longitudinal perspective on llm data contamination. In *The Twelfth International Conference on Learning Representations*.

Bettina Schlager and Steven K Feiner. 2024. Designing non-humanoid virtual assistants for task-oriented ar environments. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pages 1017–1018. IEEE.

Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. 2022. Dc-check: A data-centric ai checklist to guide the development of reliable machine learning systems. *arXiv preprint arXiv:2211.05764*.

Archit Sharma, Sedrick Keh, Eric Mitchell, Chelsea Finn, Kushal Arora, and Thomas Kollar. 2024. A critical evaluation of ai feedback for aligning large language models. *arXiv preprint arXiv:2402.12366*.

Min-Han Shih, Ho-Lam Chung, Yu-Chi Pai, Ming-Hao Hsu, Guan-Ting Lin, Shang-Wen Li, and Hung-yi Lee. 2024. Gsqa: An end-to-end model for generative spoken question answering. *Interspeech 2024*.

Dan Su and Pascale Fung. 2020. Improving spoken question answering using contextualized word representation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8004–8008.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, MA Zejun, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. In *12th International Conference on Learning Representations, ICLR 2024*.

Luis Von Ahn. 2013. Duolingo: learn a language for free while helping to translate the web. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 1–2.

Tao Wang et al. 2020. Spoken content and voice factorization for few-shot speaker adaptation. In *INTERSPEECH*, pages 796–800.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Qingyang Wu, James Gung, Raphael Shu, and Yi Zhang. 2023a. Diacttod: Learning generalizable latent dialogue acts for controllable task-oriented dialogue systems. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 255–267.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*.

Zeqiu Wu, Ryu Parish, Hao Cheng, Sewon Min, Prithviraj Ammanabrolu, Mari Ostendorf, and Hannaneh Hajishirzi. 2023b. Inscit: Information-seeking conversations with mixed-initiative interactions. *Transactions of the Association for Computational Linguistics*, 11:453–468.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Mu Yang, Bowen Shi, Matthew Le, Wei-Ning Hsu, and Andros Tjandra. 2024b. Audiobox tta-rag: Improving zero-shot and few-shot text-to-audio with retrieval-augmented generation. *arXiv preprint arXiv:2411.05141*.

Chenyu You, Nuo Chen, Fenglin Liu, Shen Ge, Xian Wu, and Yuexian Zou. 2022. End-to-end spoken conversational question answering: Task, dataset and model. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1219–1232, Seattle, United States. Association for Computational Linguistics.

Nan Yu, Guohong Fu, and Min Zhang. 2022. Speaker-aware discourse parsing on multi-party dialogues. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5372–5382.

Xiao Yu, Maximillian Chen, and Zhou Yu. 2023. Prompt-based monte-carlo tree search for goal-oriented dialogue policy planning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7101–7125.

Xiao Yu, Qingyang Wu, Yu Li, and Zhou Yu. 2024. Lions: An empirically optimized approach to align language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8732–8753.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2024. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. In *The Twelfth International Conference on Learning Representations*.

A Additional Related Work

Spoken question answering is a fundamental skill for intelligent spoken conversational agents (Khatri et al., 2018a; Zheng et al., 2024). Many tasks have been proposed in order to measure models' ability to understand spoken context (Li et al., 2018; Shih et al., 2024) and spoken requests (Faisal et al., 2021). There have also been recent works which propose newer SQA benchmarks in the post-LLM era such as Nachmani et al. (2024), but the ones there do not have a training set and so they are not directly applicable to customization approaches like DAMSEL. Previously, most approaches to SQA focused on span prediction using "cascade" approaches which include an intermediate step invoking an Automatic Speech Recognition (ASR) module followed by a fine-tuned a text classification model Chuang et al. (2020); Li et al. (2018); Su and Fung (2020) like BERT (Devlin, 2018). It is increasingly desirable to develop end-to-end pipelines to solve SQA tasks (Shih et al., 2024), particularly with the rise of generalist MLLMs (Wu et al., 2024). Such end-to-end models are desirable in speech as they afford opportunities to directly encode useful information in acoustic signals such as speaking rate, pitch, or emotions. In our work, we focus on improving methods for end-to-end SQA using both closed-weight and open-weight MLLMs.

Mixed-initiative conversations require each interlocutor to control the interaction flow (Horvitz, 1999) through the use of various pragmatic actions (Chen et al., 2023) such as clarifying questions, which can lead to to better goal completion outcomes (Guo et al., 2021; Min et al., 2020; Wu et al., 2023b). Many works focus on planning these explicit pragmatic actions (Deng et al., 2024; Yu et al., 2023), whereas other works focus on implicit (Chen et al., 2025) and continuous space actions (Wu et al., 2023a), and end-to-end generation capabilities in such settings (Li et al., 2020; Deng et al., 2022). While there have been recent efforts in designing multi-turn SQA corpora (You et al., 2022), to our knowledge, there is not yet any mixed-initiative conversation environment for speech, despite there being many additional acoustic features which may introduce ambiguity (Kurata et al., 2011; Mulholland et al., 2016). In our work, we develop the first-ever conversational SQA corpus which requires the ability to disambiguate requests and reason about clarification questions.

Adapting models with limited speech data has received much attention due to well-known problem of speaker overfitting across a variety of tasks ranging from grammatical error correction (Wang et al., 2020) to speaker verification (Jung et al., 2018). This problem is frequently addressed with the assistance of multi-task learning (Caruana, 1997). Pironkov et al. (2016) proposed a multi-task objective in which they simultaneously train a network for both ASR (their downstream task) and speaker classification. Chen and Yu (2023) found large downstream task performance improvements on speech classification tasks following a stage of multi-task pre-finetuning. In our work, we view multi-task learning through a data-centric lens. While multi-task pre-finetuning relies on additional datasets (Aghajanyan et al., 2021; Padmakumar et al., 2022), our approach improves the utilization of a fixed set of speech recordings by introducing auxiliary tasks designed to improve the cross-modal understanding capabilities of MLLMs.

B Additional Details on ASK-QA

Here, we provide several additional details on ASK-QA. First we describe the construction of the dataset in detail. Then, we demonstrate the quality of our corpus quantitatively as well as qualitatively. Specifically, we find that the audio quality is very high with strong faithfulness to the original transcript.

B.1 Dataset Construction

Overview As mentioned in Section 3, ASK-QA is constructed using Abg-CoQA as a starting point (Guo et al., 2021). Abg-CoQA is a textual conversational QA task, but as it is span-based, it does not provide very natural dialogue. Each instance consists of a passage which serves as some necessary contextual knowledge, and each conversation consists of dialogue turns where a user asks questions and an assistant is supposed to provide the correct answer or ask a clarifying question if the user's question is contextually ambiguous. The first step we take is to paraphrase each question using Gemini 1.5 Pro to convert the task into free-form QA generation.

Setting speaker roles Each written conversation can be considered a machine reading comprehension task. We break them down into three components: a story, the set of user questions, and the set of corresponding assistant responses. Our goal

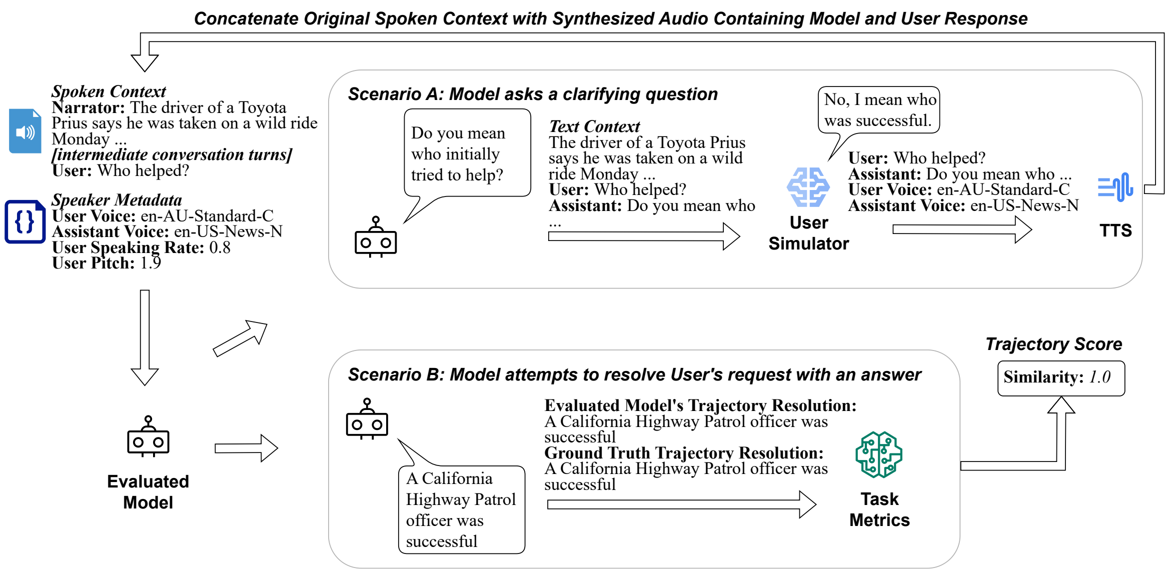

Figure A5: **Multi-turn evaluation pipeline for ASK-QA.** A model is given an audio recording containing the spoken story and spoken conversation. It is tasked with providing the correct response. While the model response is a clarifying question (as determined by a prompted Action Classifier), the model-generated response is appended to a textual version of the conversation history and shown to a user simulator. The user simulator provides a coherent response to the clarifying question, and these two generated turns are synthesized using TTS to create a new spoken context. This process repeats until the model response is not a clarifying question.

is to convert this into a listening comprehension task with two speakers having a conversation about some spoken context. Thus, for each conversation, we construct three unique speaker profiles to represent a story narrator, a user, and an assistant.

Speaker profiles Earlier works (Li et al., 2018; You et al., 2022) used commercial text-to-speech (TTS) software to synthesize speech, but at the time there were relatively high word error rate (WER) with limited options for customization. As a result, such corpora only feature a single synthetic voice without varied acoustic features (e.g. speaking rate, pitch). Here, we construct a much more diverse corpus using modern TTS solution from Google Cloud[3]. To create user speaker profiles, we aim to maximize diversity and thus sample from 38 unique voice types spanning four different accents from English-speaking countries (US, AU, GB, IN). We also randomly sample user speaking rates and pitches from a truncated normal distribution. The mean of each is set to the default value of the API endpoint. For the assistant and narrator speaker profiles, we aim to create professional-sounding dialogue, instead sampling from 26 different "news" and "studio" voices.

[3]https://cloud.google.com/text-to-speech

Text-to-Speech pipeline As per Figure 2, we then apply TTS to synthesize the story and each dialogue turn sequentially, using the appropriate speaker profile. We then concatenate the resulting audio files into a single spoken conversation. We do not adjust the default speaking rates and pitches. Following the suggestions of earlier work in text-based data synthesis (Chen et al., 2022a; Kim et al., 2023), we apply weakly supervised filtering to ensure that the synthesized speech is high quality. If any synthesized speech exceeds a WER of 0.20 (as determined by Whisper-Large v3; Radford et al. (2023)), we retry the synthesis process. If it fails three times, we discard the conversation sample. We finally randomly insert white noise into the audio by drawing from a Gaussian distribution (with an average signal to noise ratio of 21.75dB). The result is a unique speech CQA dataset with disfluencies, multiple speakers, and long audio context. The contributions of ASK-QA compared to other existing SQA datasets are in Table 1.

B.2 Additional Evaluation Details

Single-turn evaluation: We follow the standard single-turn evaluation setting with pre-determined inputs similar to existing conversational QA tasks Guo et al. (2021); Deng et al. (2022). For an evaluation instance, an agent must produce a correct response conditioned on the speech recording. In ASK-QA, the speech recording contains both the knowledge context and a multi-turn conversation context. We compare the generated answer against the ground truth response.

Multi-turn evaluation: Chen et al. (2025) propose an automatic multi-turn evaluation for Abg-CoQA, in which an agent dynamically interacts with a user simulator to work towards the goal of a conversation. Inspired by this idea, we design the first ever dynamic *speech* evaluation, which is summarized in Figure A5.

The evaluation starts by providing the evaluated MLLM with the audio file (see "Spoken Context" in Figure A5) and the text instruction. Next, an action classifier determines whether the agent response is a clarifying question or a direct answer. The action classifier is an LLM which is directly prompted with the textual conversation context, the candidate response, and the action candidates (CLARIFY/ANSWER). In this work, we use Gemini 1.5 Pro. If it is a clarifying question, we simulate a user-side response to the clarifying question using a user simulator. This user simulator is a high-capacity LLM (here, Gemini 1.5 Pro as well) provided with the textual conversation context, and a summarization of the user's ground truth information-seeking intent. Our implementation of these is the same as for Abg-CoQA in Chen et al. (2025).

In the text setting, they concatenate the original context with the newly simulated trajectory and provides this as a new prompt to the evaluated agent. Here, we apply the same TTS process as in Section 3.1 using the same speaker profiles for this conversation to dynamically create another fully spoken input. We then sample another response from the evaluated agent until a direct answer attempt is provided, or the maximum rollout depth is reached (here, we set this to three clarification turns). We apply the evaluation metric to the final response.

B.3 Conversation Examples

In Table A4, we provide a simple example of how a conversation in ASK-QA is paraphrased from the original conversation in Abg-CoQA (with the passage context omitted). One turn is grammaically incorrect and so it is paraphrased using Gemini 1.5 Pro.

In Table A5, we provide an example of a conversation in ASK-QA with more perturbations from Abg-CoQA. Turn 4 is rephrased as a complete sen-

Version	Turn No.	Speaker	Content
Abg-CoQA	1	User	what was the others reaction?
	2	Assistant	Everyone was happy
	3	User	was he hungry?
	4	Assistant	Do you mean at first?
	5a	User	Yes
	6a	Assistant	Yes
	5b	User	No, I mean ultimately.
	6b	Assistant	No, he was not hungry anymore.
ASK-QA	1	User	what was the others reaction?
	2	Assistant	Everyone was happy.
	3	User	was he hungry?
	4	Assistant	Do you mean at first?
	5a	User	Yes
	6a	Assistant	**Yes, he was hungry at first.**
	5b	User	No, I mean ultimately.
	6b	Assistant	No, he was not hungry anymore.

Table A4: **Comparison of the dialogue turns from a conversation in Abg-CoQA and the paraphrased version in ASK-QA.** Bold: text paraphrased using Gemini. We do not perturb the user turns in order to maintain the original linguistic diversity. However, for the Assistant turns, we paraphrase the language concisely if necessary in order to ensure that the speech is grammatically correct. Here, the only dialogue turn that differs is 6a. (a) and (b) denote differing trajectories, which are the turns that the Assistant has to navigate successfully during evaluation.

Version	Turn No.	Speaker	Content
Abg-CoQA	1	User	Are they related?
	2	Assistant	yes
	3	User	How?
	4	Assistant	brothers
	5	User	Where do they put the lemonade stand?
	6	Assistant	by the sidewalk
ASK-QA	1	User	Are they related?
	2	Assistant	Yes, they are related.
	3	User	How?
	4	Assistant	**They are brothers.**
	5	User	Where do they put the *the* lemonade stand ?
	6	Assistant	**They put the lemonade stand by the sidewalk.**

Table A5: **A modified conversation in ASK-QA.** Bold: paraphrased text using Gemini. Italics: repeat disfluency injected using LARD (Passali et al., 2022).

tence. Turn 5 injects a repeat disfluency into the User-side speech. Turn 6 is also rephrased as a complete sentence.

Table A11 provides a full example of a full example from the ASK-QA dataset. We include the passage context, as well as the provided dialogue excerpt. We denote the input modalities as well as our instruction for response generation using the MLLM.

Our supplementary material contains an example of an evaluation instance in our dataset.

C Efficient Multimodal Adapters via Audio Representation Projection

As demonstrated in Section 3, our data-centric approach is easily applicable to both settings with access to tuning APIs for closed-source MLLMs like Gemini, and settings with access to open-weight models for each modality. Here, we describe our approach in the open-weight scenario.

Textual instructions serve as a highly controllable interface, and as such, recent work has found much success in unifying multiple modalities with large pre-trained decoder-only language models (Liu et al., 2024; Arora et al., 2024; Kong et al., 2024). These works aim to leverage the impressive instruction-following capabilities of LLMs to interpret additional modalities (e.g. vision, speech, video etc.) by effectively mapping their representations to LLM input space.

Architecture: In our work, we consider the high-level architecture presented in Ma et al. (2024). We projecting the speech input represented by an audio encoder into the embedding space of an LLM to improve performance on ASR tasks, only tuning the weights of a linear projection layer and freezing the other model components.[4] As described in Section 3, our speech encoder is WavLM-Large (Chen et al., 2022b). We primarily experimented with tuning Qwen 2.5-Instruct (Yang et al., 2024a) with 7B parameters as our base decoder-only LLM. We also experimented with Phi 3.5 Mini (Abdin et al., 2024) with 3B parameters in Table A7. These MLLMs are referred to as Speech-Qwen and Speech-Phi, respectively. We tune this adapter using standard cross-entropy loss. Details on our tuning experiments are provided in Appendix C.

[4]As in Ma et al. (2024), the projection layer consists of merely 17.8M parameters for the proposed models.

Projection Layer Pre-training: While this projection layer is tuned directly on the target ASR task in Ma et al. (2024), we find that this approach may struggle with direct single-task fine-tuning on our more difficult SQA tasks which do not have the same abundance of data. Similarly to how visual MLLMs are often pre-trained on image captioning (Liu et al., 2024), we pretrain the projection layer for one epoch on large-scale ASR data.

D Additional Details on Data-Centric Multi-Task Learning

Figure A6 provides a high-level overview of our multi-task learning approach. On the left, we show examples of each of our SQA corpora used for experimentation. At a high level, each corpus consists of passage and a conversation. In ASK-QA, the contextual inputs are fully spoken. In Spoken-SQuAD, the knowledge is spoken while the question is written. In SD-QA, the knowledge is written while the question is spoken (in multiple languages and regional dialects).

Regardless of the input modalities, each instance can be mapped to new data instances representing the auxiliary tasks in Section 2. The visible examples on the right side of Figure A6 are our multi-task instances for Spoken-SQuAD. The top-right panel is our Listening Comprehension task, our middle-right panel is our Cross-Modal Commonsense Reasoning task, and our bottom right task is the standard QA task (which is just reorganized from the middle-left panel).

D.1 Multi-Task Training Examples for ASK-QA

We provide concrete examples of the auxiliary tasks for a single instance of ASK-QA in Tables A8, A9, A10, A11. Each of these tables has the exact same speech recording. Table A8 demonstrates how the ground-truth answer is joined to negatively sampled answers from other QA pairings to form the response selection task. Table A11 is similar and demonstrates the textual instruction used to steer the MLLM to directly generate the ground-truth answer. As stated in Section 3.1, we break the Listening Comprehension task into two components since each recording comprises a narrated story and a conversation. Table A9 demonstrates steering the MLLM to transcribe the conversation. Table A10 demonstrates steering the MLLM to transcribe the narrated story.

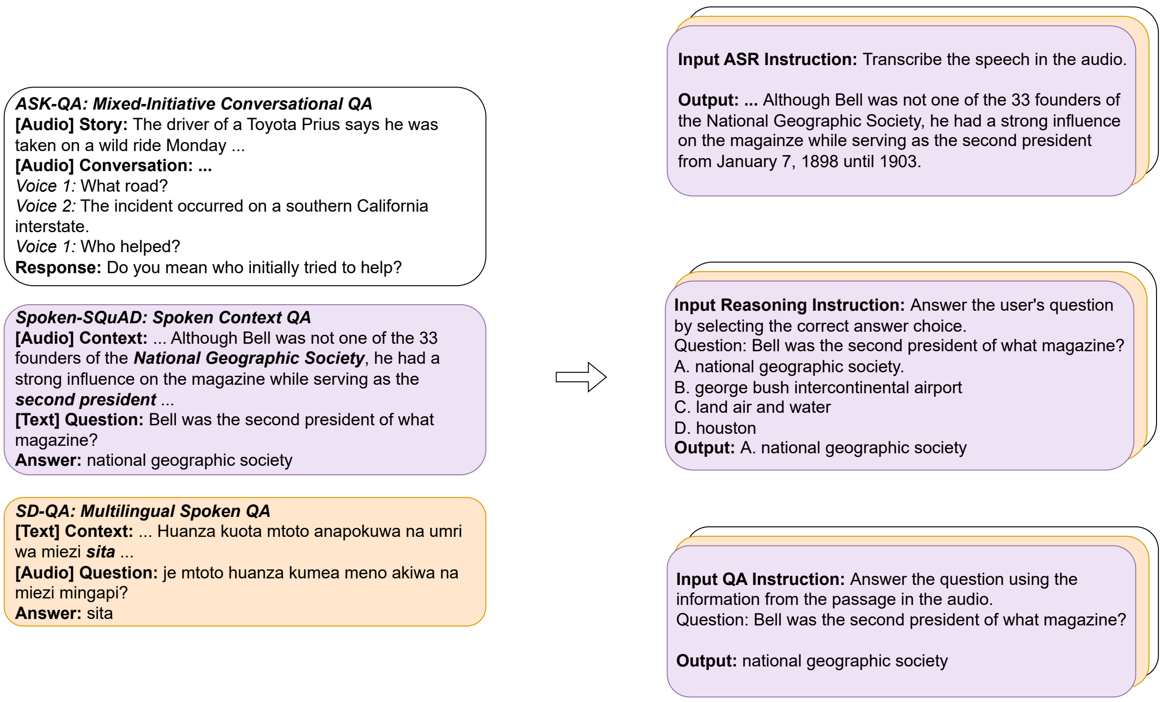

Figure A6: **Creating multi-task data from individual SQA training instances.** Left: examples of instance metadata from the three SQA datasets used in this paper. Right: for each speech-QA pairing, we are able to form three tasks designed to teach MLLMs' cross-modal reasoning ability.

E Extended Experimental Results

Due to space constraints in the main text, we describe our additional experiments in this section. In particular, we apply DAMSEL to a third corpus which features multi-lingual SQA scenarios, and we look at additional baseline comparisons and base MLLMs.

E.1 Additional Experiments on ASK-QA

Our main findings and results are presented in Section 3.1. Here, we present our full results on ASK-QA. Specifically, we additionally examine the efficacy of our approach with an additional closed-source MLLM, Gemini Flash. Our results in Figure 3 also highlight trajectory-level similarity, and here, we also present results on single-turn evaluation.

Base Model	Approach	Data	Single-Turn Sim. ↑	Multi-Turn Sim. ↑
Gemini Flash	Prompt	0%	65.10	64.45
Gemini Pro	Prompt	0%	63.20	62.85
Gemini Pro	Single	1%	74.10	72.29
Gemini Pro	DAMSEL	1%	**77.64**	**76.66**
Gemini Pro	Single	10%	**75.82**	**74.60**
Gemini Pro	DAMSEL	10%	**79.13**	**77.62**
Gemini Pro	Single	100%	80.26	78.85
Gemini Pro	DAMSEL	100%	**81.40**	**80.12**
Gemini Flash	Single	1%	70.43	70.60
Gemini Flash	DAMSEL	1%	**73.88**	**73.01**
Gemini Flash	Single	10%	76.21	74.89
Gemini Flash	DAMSEL	10%	**77.38**	**75.49**
Gemini Flash	Single	100%	79.10	77.94
Gemini Flash	DAMSEL	100%	**80.47**	**79.30**
Speech-Qwen	Single	1%	47.63	47.31
Speech-Qwen	DAMSEL	1%	**54.54**	**53.60**
Speech-Qwen	Single	10%	63.43	62.71
Speech-Qwen	DAMSEL	10%	**68.27**	**67.58**
Speech-Qwen	Single	100%	69.63	68.80
Speech-Qwen	DAMSEL	100%	**71.85**	**71.09**

Table A6: **Comparing single-task tuning to DAMSEL, our multi-task fine-tuning approach, on ASK-QA's test set**.

E.2 Additional Experiments on Spoken-SQuAD

In Table A7, we provide our extended results on the Spoken-SQuAD corpus.

Additional Models and Baselines: We additionally examine experiments with Speech-Phi, which we train as described in Appendix C. This model uses Phi 3.5 Mini as the base decoder-only LLM, with up to 128k context.

We also provide the full experimental results of several baselines: FusionNet from Huang (2017), QANet from Lee et al. (2019), DDNet which is the state-of-the-art open-source model from You et al. (2022), and Whisper-Qwen, which is a

Base Model	Approach	Data	EM ↑	F1 ↑
FusionNet (Huang, 2017)	–	100%	46.51	60.06
QANet (Lee et al., 2019)	–	100%	49.60	61.85
DDNet (You et al., 2022)	–	100%	64.10	77.10
Whisper-Qwen	Prompt	0%	59.13	74.08
Whisper-Qwen	Prompt	20-shot	70.00	79.50
Gemini Pro	Prompt	0%	67.41	82.21
Speech-Phi	Single	1%	15.08	25.03
Speech-Phi	DAMSEL	1%	**22.91**	**35.02**
Speech-Phi	Single	10%	31.43	44.69
Speech-Phi	DAMSEL	10%	**49.32**	**63.09**
Speech-Phi	Single	100%	50.53	64.46
Speech-Phi	DAMSEL	100%	**62.14**	**74.31**
Speech-Qwen	Single	1%	60.25	73.24
Speech-Qwen	DAMSEL	1%	**63.15**	**75.40**
Speech-Qwen	Single	10%	62.69	75.94
Speech-Qwen	DAMSEL	10%	**66.38**	**78.80**
Speech-Qwen	Single	100%	68.75	80.92
Speech-Qwen	DAMSEL	100%	**72.13**	**82.36**

Table A7: **Experimental results comparing single-task SFT (ST) and our proposed multi-task approach (MT) on Spoken SQuAD's test set**.

cascade-style system which uses uses Whisper-Large v3 (Radford et al., 2023) to first transcribe the audio then passes the transcription as context to Qwen 2.5 7B Instruct (Yang et al., 2024a) (the same model used for tuning in our experiments). We use this modular Whisper-Qwen system with both 0-shot prompting and 20-shot in-context learning. The in-context examples are given using fully textual gold transcription examples.

Findings: In Table A7, we consistently see that in the end-to-end speech setting, multi-task learning improves upon single-task learning. We see a particularly strong improvement using Speech-Phi. We also note that the final ability of the adapter-trained MLLM to complete the downstream SQA task may depend on the base decoder's performance on textual QA. If the projection layer is tuned to perfectly represent the audio, then the bottleneck on performance may be the decoder model's task performance on SQuAD since Spoken-SQuAD is a fully semantic task with limited acoustic diversity – the focus in the corpus construction at the time was on discrepancies between TTS and ASR (Li et al., 2018). We see that providing Qwen with golden transcripts for in-context learning in a modular system can achieve very strong performance for this very reason.

F Training Details

Open-weight models: Our tuning experiments using open-weight models are conducted on a single node with 8 NVIDIA A100 80GB GPUs. We rely on Deepspeed ZeRO-3 (Rasley et al., 2020) and build on top of HuggingFace (Wolf et al., 2020), PyTorch (Paszke et al., 2019), and SLAM-LLM (Ma et al., 2024). For both Speech-Qwen and Speech-Phi, we achieve our best results using an initial learning rate of 1e-4. With Speech-Qwen, we use a total batch size of 8 given our hardware constraints. For Speech-Phi, the total batch size is 16, 16, and 32 for ASK-QA, Spoken-SQuAD, and SD-QA, respectively. Our models are tuned on downstream tasks for up to 20 epochs in the limited data setting, with early stopping based on validation loss.

Closed-weight models: We perform supervised fine-tuning on "gemini-1.5-flash-002" and "gemini-1.5-pro-002" using adapters on Google Cloud's Vertex AI platform. We obtain best results using a learning rate multiplier of 1. We tune our models for a maximum of 20 epochs in the limited data setting.

G Risks and Ethical Considerations

There are significant privacy concerns around speech data collection (Nautsch et al., 2019), and so in this work, we rely on synthetically generated speech. However, as previously mentioned, one limitation of our work is on TTS quality. It is possible that generating long-context speech at scale will allow for hallucinations depending on the quality of the chosen TTS model. Even with automated filtering efforts, it may still be possible for these hallucinations to bypass the filtering mechanism. In our corpus, the Word Error Rate should be rather low due to the aforementioned filtering mechanism, but this still poses risk – especially if such synthetic data are contributed to large-scale model training.

H Artifacts Used

I Assets Used

All resources used have been cited appropriately in the paper. In this section, we enumerate each of the existing artifacts used in our work along with their license.

Existing Models

- Gemini 1.5 Pro (gemini-1.5-pro-002), Gemini 1.5 Flash (gemini-1.5-flash-002) (Gemini Team et al., 2023): Accessed through the Google Cloud Vertex AI Platform. https://cloud.google.com/products/gemini?hl=en

- MiniLM-L6-v2 (Reimers, 2019): Apache 2.0. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

- Qwen2.5-7B-Instruct (Yang et al., 2024a): MIT Open-Source License. https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

- WavLM (Chen et al., 2022b): MIT Open-Source License. https://github.com/microsoft/unilm/blob/master/wavlm/README.md

- Phi-3-mini-128k-instruct (Abdin et al., 2024): MIT Open-Source License. https://huggingface.co/microsoft/Phi-3-mini-128k-instruct

Existing Datasets

- Abg-CoQA (Guo et al., 2021): MIT Open-Source License. https://github.com/MeiqiGuo/AKBC2021-Abg-CoQA

- Spoken-SQuAD (Li et al., 2018): Open-Source. https://github.com/Chia-Hsuan-Lee/Spoken-SQuAD

- SQuAD (Rajpurkar et al., 2016): CC-BY-SA 4.0 License. https://rajpurkar.github.io/SQuAD-explorer/

Existing and Software

- Google Cloud Pipeline Components: Apache 2.0. https://cloud.google.com/vertex-ai/docs/pipelines/components-introduction

- HuggingFace Transformers (?): Apache 2.0. https://github.com/huggingface/transformers/tree/main

- PyTorch (Paszke et al., 2019): PyTorch Open Source License. https://github.com/pytorch/pytorch/tree/main

- Vertex AI SDK: Apache 2.0. https://cloud.google.com/vertex-ai/docs/python-sdk/use-vertex-ai-python-sdk

- SLAM-LLM: MIT License. https://github.com/X-LANCE/SLAM-LLM/tree/main

Usage	Modality	Content
Input	Speech	**Speaker 1:** A few years ago, an Englishman called Roy Jones went on holiday to a small seaside town in the west of England. He was swimming in the sea one day when, as he opened his mouth, his false teeth fell out and floated away. The following year, Mr. Jones returned to the same town. As he was having dinner in a local cafe one evening, he mentioned the story of his lost teeth to the manager. The manager looked surprised. He explained that he had found a set of false teeth on the beach last month. Then he asked Roy Jones if he wanted to try them on. OK, said Mr. Jones. I suppose it won't do any harm. When the manager brought him the teeth, Mr. Jones put them into his mouth, and laughed and laughed. They were his. In 1987, an American couple called Jane and Robert Bentley went for a picnic on a beach in California. **When they returned home, Mrs. Bentley realized that she had lost her wedding ring. It wasn't a lot of money but it was valuable to Jane Bentley. The Bentleys drove straight back to the beach, and searched for the ring for three hours, but could not find it. A few months later, Mr. Bentley went fishing off the same beach. As he pulled a large crab out of the sea, he noticed that there was something attached to one of its claws. It was his wife's wedding ring!** At the end of the 19th century, a young woman called Rose Harcourt was on her honeymoon in Barmouth, North Wales, when she lost a gold bracelet her husband had given her as a wedding gift. Feeling very upset, she went straight to the police stations and asked if anyone had found her bracelet. Unfortunately, no one had. Twenty-five years later, the Harcourts returned to Barmouth. They were sitting on the beach one day when Mrs. Harcourt noticed something gold in the sand by the edge of the sea. She walked down to see what it was, and discovered her gold bracelet that had been missing for 25 years. **Speaker 2:** Was it expensive? **Speaker 3:** No, it was not expensive. **Speaker 2:** Was it recovered? **Speaker 3:** Yes, it was recovered. **Speaker 2:** When?
Input	Text	The audio recording contains a story followed by a conversation between a User and an Assistant. You will continue the conversation for the Assistant by selecting the most appropriate response from the following: A. Do you mean the popular generic name? B. Are you asking why the dog was looking at Sue or why Jack walked up to Sue? C. More Chinese people can afford cars because of them. **D. It was recovered a few months later.**
Output	Text	**D. It was recovered a few months later.**

Table A8: Example of the commonsense Response Selection auxiliary task for ASK-QA.

Usage	Modality	Content
Input	Speech	**Speaker 1:** A few years ago, an Englishman called Roy Jones went on holiday to a small seaside town in the west of England. He was swimming in the sea one day when, as he opened his mouth, his false teeth fell out and floated away. The following year, Mr. Jones returned to the same town. As he was having dinner in a local cafe one evening, he mentioned the story of his lost teeth to the manager. The manager looked surprised. He explained that he had found a set of false teeth on the beach last month. Then he asked Roy Jones if he wanted to try them on. OK, said Mr. Jones. I suppose it won't do any harm. When the manager brought him the teeth, Mr. Jones put them into his mouth, and laughed and laughed. They were his. In 1987, an American couple called Jane and Robert Bentley went for a picnic on a beach in California. When they returned home, Mrs. Bentley realized that she had lost her wedding ring. It wasn't a lot of money but it was valuable to Jane Bentley. The Bentleys drove straight back to the beach, and searched for the ring for three hours, but could not find it. A few months later, Mr. Bentley went fishing off the same beach. As he pulled a large crab out of the sea, he noticed that there was something attached to one of its claws. It was his wife's wedding ring! At the end of the 19th century, a young woman called Rose Harcourt was on her honeymoon in Barmouth, North Wales, when she lost a gold bracelet her husband had given her as a wedding gift. Feeling very upset, she went straight to the police stations and asked if anyone had found her bracelet. Unfortunately, no one had. Twenty-five years later, the Harcourts returned to Barmouth. They were sitting on the beach one day when Mrs. Harcourt noticed something gold in the sand by the edge of the sea. She walked down to see what it was, and discovered her gold bracelet that had been missing for 25 years. **Speaker 2:** Was it expensive? **Speaker 3:** No, it was not expensive. **Speaker 2:** Was it recovered? **Speaker 3:** Yes, it was recovered. **Speaker 2:** When?
Input	Text	The audio recording contains a story followed by a conversation between a User and an Assistant. Transcribe the conversation but not the story. Provide your answer in the format User: [Utterance] Assistant: [Utterance] and so on.
Output	Text	**User:** Was it expensive? **Assistant:** No, it was not expensive. **User:** Was it recovered? **Assistant:** Yes, it was recovered. **User:** When?

Table A9: Example of the Dialogue Listening Comprehension auxiliary task for ASK-QA.

Usage	Modality	Content
Input	Speech	**Speaker 1:** A few years ago, an Englishman called Roy Jones went on holiday to a small seaside town in the west of England. He was swimming in the sea one day when, as he opened his mouth, his false teeth fell out and floated away. The following year, Mr. Jones returned to the same town. As he was having dinner in a local cafe one evening, he mentioned the story of his lost teeth to the manager. The manager looked surprised. He explained that he had found a set of false teeth on the beach last month. Then he asked Roy Jones if he wanted to try them on. OK, said Mr. Jones. I suppose it won't do any harm. When the manager brought him the teeth, Mr. Jones put them into his mouth, and laughed and laughed. They were his. In 1987, an American couple called Jane and Robert Bentley went for a picnic on a beach in California. When they returned home, Mrs. Bentley realized that she had lost her wedding ring. It wasn't a lot of money but it was valuable to Jane Bentley. The Bentleys drove straight back to the beach, and searched for the ring for three hours, but could not find it. A few months later, Mr. Bentley went fishing off the same beach. As he pulled a large crab out of the sea, he noticed that there was something attached to one of its claws. It was his wife's wedding ring! At the end of the 19th century, a young woman called Rose Harcourt was on her honeymoon in Barmouth, North Wales, when she lost a gold bracelet her husband had given her as a wedding gift. Feeling very upset, she went straight to the police stations and asked if anyone had found her bracelet. Unfortunately, no one had. Twenty-five years later, the Harcourts returned to Barmouth. They were sitting on the beach one day when Mrs. Harcourt noticed something gold in the sand by the edge of the sea. She walked down to see what it was, and discovered her gold bracelet that had been missing for 25 years. **Speaker 2:** Was it expensive? **Speaker 3:** No, it was not expensive. **Speaker 2:** Was it recovered? **Speaker 3:** Yes, it was recovered. **Speaker 2:** When?
Input	Text	The audio recording contains a story followed by a conversation between a User and an Assistant. Transcribe the story but not the conversation.
Output	Text	A few years ago, an Englishman called Roy Jones went on holiday to a small seaside town in the west of England. He was swimming in the sea one day when, as he opened his mouth, his false teeth fell out and floated away. The following year, Mr. Jones returned to the same town. As he was having dinner in a local cafe one evening, he mentioned the story of his lost teeth to the manager. The manager looked surprised. He explained that he had found a set of false teeth on the beach last month. Then he asked Roy Jones if he wanted to try them on. OK, said Mr. Jones. I suppose it won't do any harm. When the manager brought him the teeth, Mr. Jones put them into his mouth, and laughed and laughed. They were his. In 1987, an American couple called Jane and Robert Bentley went for a picnic on a beach in California. When they returned home, Mrs. Bentley realized that she had lost her wedding ring. It wasn't a lot of money but it was valuable to Jane Bentley. The Bentleys drove straight back to the beach, and searched for the ring for three hours, but could not find it. A few months later, Mr. Bentley went fishing off the same beach. As he pulled a large crab out of the sea, he noticed that there was something attached to one of its claws. It was his wife's wedding ring! At the end of the 19th century, a young woman called Rose Harcourt was on her honeymoon in Barmouth, North Wales, when she lost a gold bracelet her husband had given her as a wedding gift. Feeling very upset, she went straight to the police stations and asked if anyone had found her bracelet. Unfortunately, no one had. Twenty-five years later, the Harcourts returned to Barmouth. They were sitting on the beach one day when Mrs. Harcourt noticed something gold in the sand by the edge of the sea. She walked down to see what it was, and discovered her gold bracelet that had been missing for 25 years.

Table A10: Example of the Story Listening Comprehension auxiliary task for ASK-QA.

Usage	Modality	Content
Input	Speech	**Speaker 1:** A few years ago, an Englishman called Roy Jones went on holiday to a small seaside town in the west of England. He was swimming in the sea one day when, as he opened his mouth, his false teeth fell out and floated away. The following year, Mr. Jones returned to the same town. As he was having dinner in a local cafe one evening, he mentioned the story of his lost teeth to the manager. The manager looked surprised. He explained that he had found a set of false teeth on the beach last month. Then he asked Roy Jones if he wanted to try them on. OK, said Mr. Jones. I suppose it won't do any harm. When the manager brought him the teeth, Mr. Jones put them into his mouth, and laughed and laughed. They were his. In 1987, an American couple called Jane and Robert Bentley went for a picnic on a beach in California. **When they returned home, Mrs. Bentley realized that she had lost her wedding ring. It wasn't a lot of money but it was valuable to Jane Bentley. The Bentleys drove straight back to the beach, and searched for the ring for three hours, but could not find it. A few months later, Mr. Bentley went fishing off the same beach. As he pulled a large crab out of the sea, he noticed that there was something attached to one of its claws. It was his wife's wedding ring!** At the end of the 19th century, a young woman called Rose Harcourt was on her honeymoon in Barmouth, North Wales, when she lost a gold bracelet her husband had given her as a wedding gift. Feeling very upset, she went straight to the police stations and asked if anyone had found her bracelet. Unfortunately, no one had. Twenty-five years later, the Harcourts returned to Barmouth. They were sitting on the beach one day when Mrs. Harcourt noticed something gold in the sand by the edge of the sea. She walked down to see what it was, and discovered her gold bracelet that had been missing for 25 years. **Speaker 2:** Was it expensive? **Speaker 3:** No, it was not expensive. **Speaker 2:** Was it recovered? **Speaker 3:** Yes, it was recovered. **Speaker 2:** When?
Input	Text	The audio recording contains a story followed by a conversation between a User and an Assistant. You will continue the conversation for the Assistant by providing the next response. If you do not have enough information or context to answer the User's last question, you should ask a clarifying question.
Output	Text	It was recovered a few months later.

Table A11: Example of the Response Generation task for ASK-QA.

Usage	Modality	Content
Input	Speech	Alexander Graham Bell. March 3. 1847 August 2. 1922. Was a Scottish-born. N3. Scientist. Inventor. Engineer and innovator who is credited with patenting the first practical telephone.
Input	Text	The audio contains a passage providing some information. Follow the User's request. User: Transcribe the speech in the passage.
Output	Text	Alexander Graham Bell. March 3. 1847 August 2. 1922. Was a Scottish-born. N3. Scientist. Inventor. Engineer and innovator who is credited with patenting the first practical telephone.

Table A12: Example of the Listening Comprehension auxiliary task for Spoken-SQuAD.

Usage	Modality	Content
Input	Speech	Alexander Graham Bell. March 3. 1847 August 2. 1922. Was a Scottish-born. N3. Scientist. Inventor. Engineer and innovator who is credited with patenting the first practical **telephone**.
Input	Text	The audio contains a passage providing some information. The user will ask a question about some information from the audio. The assistant should answer the user's question by selecting the correct answer choice. User: What is Bell most famous for inventing? Choose from the following choices: A. britain **B. telephone** C. major performing arts D. london county council
Output	Text	**B. telephone**

Table A13: Example of the commonsense Response Selection auxiliary task for Spoken-SQuAD.

Usage	Modality	Content
Input	Speech	Alexander Graham Bell. March 3. 1847 August 2. 1922. Was a Scottish-born. N3. Scientist. Inventor. Engineer and innovator who is credited with patenting the first practical **telephone**.
Input	Text	The audio contains a passage providing some information. The user will ask a question about some information from the audio. The assistant should answer the user's question using information which can be found in the passage. User: What is Bell most famous for inventing?
Output	Text	**telephone**

Table A14: Example of the Response Generation auxiliary task for Spoken-SQuAD.