



Approaching Dialogue State Tracking via Aligning Speech Encoders and LLMs

Šimon Sedláček¹, Bolaji Yusuf¹, Ján Švec¹, Pradyoth Hegde^{1,2}, Santosh Kesiraju¹, Oldřich Plchot¹, Jan Černocký¹

¹Speech@FIT, Brno University of Technology, Czechia

²Indian Institute of Information Technology Dharwad, India

{isedlacek, iyusuf, isvecjan, kesiraju, iplchot, cernocky}@fit.vut.cz,
pradyothhegde@gmail.com

Abstract

In this work, we approach spoken Dialogue State Tracking (DST) by bridging the representation spaces of speech encoders and LLMs via a small connector module, with a focus on fully open-sourced and open-data components (WavLM-large, OLMo). We focus on ablating different aspects of such systems including full/LoRA adapter fine-tuning, the effect of agent turns in the dialogue history, as well as fuzzy matching-based output post-processing, which greatly improves performance of our systems on named entities in the dialogue slot values. We conduct our experiments on the SpokenWOZ dataset, and additionally utilize the Speech-Aware MultiWOZ dataset to augment our training data. Ultimately, our best-performing WavLM + connector + OLMo-1B aligned models achieve state of the art on the SpokenWOZ test set (34.66% JGA), and our system with Gemma-2-9B-instruct further surpasses this result, reaching 42.17% JGA on SpokenWOZ test.

Index Terms: dialogue state tracking, task-oriented dialogue, speech LLMs

1. Introduction

Task-oriented dialogues (ToD) are multiturn conversations between a user and an agent, where the former has a specific goal (e.g., booking a restaurant for 5 people on Friday night) that is achieved with the help of the agent. A key component of automated ToD systems is dialogue state tracking (DST), the task of tracking the user’s intent (e.g. “book-restaurant”) and identifying the slots (e.g. “restaurant-people”: “5”, “day”: “Friday”). Recently, aided by the DSTC-11 challenge [1] and the development of realistic SpokenWoZ [2] dataset, the scope of DST research is slowly advancing from being exclusively text-based to speech domain.

A typical approach to DST from spoken conversations is via a cascade of several systems: automatic speech recognition (ASR) → error correction module → text-based DST [1, 3]. Although end-to-end (E2E) DST training offers an attractive alternative due to the simplicity of training and the potential to avoid cascading errors, E2E DST models are difficult to train due to the scarcity of data compared to other speech processing tasks. The paradigm of interconnecting pre-trained speech encoders with large language models (LLMs) offers a solution to the issue and has shown performance that is competitive with cascade systems [4]. However, most prior work relies on fully or partly closed models for which either the model weights or training data are not openly available.

Modality matching is one of the key challenges in aligning pre-trained speech and language models. Various approaches such as text-representation up-sampling [5], speech-representation sub-sampling and fixed-length representations

based on Q-formers have been studied in the past [6, 7]. The application of such speech language models was not only studied on traditional tasks such as ASR, speech translation and text-to-speech [8, 9], but also on language understanding, question-answering tasks and dialogue state tracking [10, 11, 12]. Retaining the ability of pre-trained LLMs even after the alignment is one of the main challenges [4], as simple connector based approaches [13] do not generalize well beyond the domain of the training data [14]. Our work complements the prior works, as we study the task of dialogue state tracking from spoken conversations with fully open speech and language models.

In this paper, we propose an end-to-end DST model based on fully open components. We connect a WavLM speech [15] encoder with an OLMo [16] LM through a small trainable connector module, which we first train to align the representation spaces of the pre-trained modes and then fine-tune the resulting model directly for DST. We conducted experiments on the SpokenWoZ [2] and Speech-Aware MultiWoZ [1] datasets, showing that our proposed system with the fully open OLMo-1B model already yields state-of-the-art performance on SpokenWoZ and that we can obtain further significant improvements when we use the Gemma-2-9B-Instruct model.

2. Method overview

Our method treats DST as the problem of mapping from user speech and user-agent dialogue history to a dictionary containing the correct transcription of the user’s speech, and the inferred dialogue states (domains, slots). Therefore, we adopt a model which we train to take speech as input and directly output a JSON string representing this dictionary at each turn of the dialogue.

The model is composed of three parts: a pretrained speech encoder, a pretrained LLM and a small connector module that joins the speech encoder and the LLM. The encoder computes a representation of the speech input at the current turn of the conversation. Then, the speech representations are downsampled to better match the granularity of the LLM text input, and connector module maps them into the text embedding space of the LM, where they act as soft prompts. Finally, the LM autoregressively generates the tokens that comprise the JSON string. The LM input is prefixed with the connector output to provide speech conditioning and by the dialogue history in the form of the transcription of previous user inputs and agent responses.

2.1. Training

Starting from pretrained encoder, we train the model in the two stages illustrated in Figure 1: an ASR pretraining stage and a joint ASR-DST finetuning.

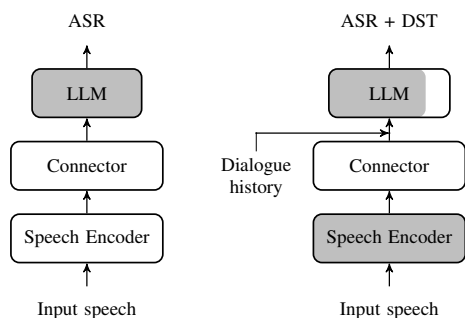


Figure 1: (Left) Stage-1 pre-training for ASR. (Right) Stage-2 training for joint ASR and dialogue state tracking (DST). Shaded modules are frozen during training. Partly shaded implies that the module is frozen, but additional trainable parameters are added via LoRA.

2.1.1. ASR pre-training

Since the speech encoder and LM are pretrained independently of each other, we first bridge their representation spaces before fine-tuning for DST. In this first stage, we freeze the LM and finetune the encoder and connector for ASR, conditioning the LLM solely using the connector outputs¹.

Thus, we are able to leverage large scale ASR datasets (which are orders of magnitude larger than typical DST datasets) to obtain robust alignment of the encoder and the LM. Moreover, ASR training is necessary since the dialogue history for the final DST model requires transcribing the user input.

2.1.2. Joint ASR-DST finetuning

In the second training stage, we keep the speech encoder frozen, and introduce LoRA [17] adapters into the LM, which we finetune along with the connector for the target DST task.

For each turn in a conversation, we append the ground-truth dialogue history to the connector output speech embeddings, and train the model to minimize the negative log-likelihood of the JSON string encoding the dialogue state for the current turn – the transcription, domains, and slots.

Note that it is imperative that we still train jointly for ASR and DST since the model will not have access to the full ground-truth dialogue history at inference time, and will have to rely on its own ASR capabilities to get the user side of the dialogue.

2.2. Inference and post-processing

Inference proceeds in a turn-by-turn fashion. For each turn, the JSON string outputted by the model is converted into the corresponding dictionary whose elements are read as the dialogue state and the dictionary field corresponding to the user ASR transcription is appended to the previous dialogue history for use in subsequent turns of the dialogue.

To obtain the final dialogue slots, we employ a fuzzy matching² scheme, which is typically used in the DST community [3]. This scheme maps the given slot value to the closest one in the database.

¹ASR prompt format: '%speech_embeds% {"transcription": %labels%}'

²pypi.org/project/fuzzywuzzy/

3. Experiments

3.1. Datasets

We conduct our experiments using two speech-grounded task-oriented dialogue datasets: SpokenWoZ [2] (SWOZ) and Speech-Aware MultiWoZ [1] (MWOZ), built on top of the original text-based MultiWoZ 2.1 [18, 19]. We remove nine originally corrupted conversations from the SpokenWoZ test set³. For SpokenWoZ, we also generate ASR transcripts using Whisper-large-v3 [20] both to serve as better reference text for training the end-to-end system, and as input to our cascaded systems. For evaluation, we adopt the standard MWOZ evaluation script [21]⁴ and report Joint Goal Accuracy (JGA) as well as Slot Error Rate (SER) for all test sets.

For the speech encoder/connector ASR pre-training stage, we use the Fisher-Switchboard (2000h) [22] Librispeech (1000h) [23] and How2 (300h) [24] datasets.

3.2. Models and training configuration

We use WavLM-large⁵ [15] pre-trained on 96k hours of speech data (Libri-Light, VoxPopuli and GigaSpeech), as the speech encoder in our experiments. We primarily use OLMo-1B [16] as the pretrained language model. We choose this model to address test set contamination concerns, as their training data are openly documented. Additionally, we conduct some experiments using Gemma-2-9B-Instruct⁶ [25, 26] as the base LM in order to facilitate comparison with some prior work.

The connector module is a 2-layer transformer encoder with 16 attention heads, hidden size 1024 and feedforward intermediate dimension of 4096. This transformer is preceded by a sub-sampling layer, which stacks 6 neighbouring WavLM embeddings and projects them into the hidden size of the connector, resulting in 6x downsampling.

In the ASR pretraining stage, we train with batch size of 64, learning rate of 1e-4 and 2000 warmup steps and train until the cross-entropy on the ASR dev sets stops improving.

In the joint ASR-DST training stage, we set the rank of the LoRA layers to r=16, and train with batch size of 128, learning rate of 5e-5 and 500 warmup steps on both SWOZ and MWOZ until the cross-entropy on the combined dev set stops improving. Then, optionally, we perform one additional epoch on finetuning separately for each dataset, with a batch size of 256 on the target dataset (referred to subsequently as FT-sw for SWOZ and FT-mw for MWOZ).

3.3. Baselines and Cascaded Systems

We establish a number of baseline cascaded Whisper/OLMo-1B DST systems, experimenting with full fine-tuning of the foundation LLMs, LoRA adapters, effects of providing agent-side dialogue history, and post-processing.

When training, we compute loss over the whole dialogue prompt schema (even the history), when decoding, only the domain and slot predictions are generated. Whisper-large-v3 is used as the ASR frontend instead of the original SpokenWoZ transcripts. The best-performing cascaded system on SpokenWoZ test is shown in row three of Table 1.

³github.com/AlibabaResearch/DAMO-ConvAI/issues/87

⁴github.com/Tomiinek/MultiWOZ_Evaluation

⁵huggingface.co/microsoft/wavlm-large

⁶huggingface.co/google/gemma-2-9b-it

Table 1: *SpokenWoZ JGA comparison of our best-performing and cascaded and E2E models with two reference systems (1, 2) from prior works.*

Model	SWOZ test	
	JGA	SER
(1) Gemma-2-9B-Instruct (cascaded) [27]	25.40	-
(2) SPACE+WavLM.align [2]	25.65	-
Whisper→OLMo-1B, full FT, SW+MW (cascade)	30.74	31.11
WavLM + conn. + OLMo-1B (A11 from Table 3)	34.66	26.80
WavLM + conn. + Gemma-2-9B-Instruct (Table 4)	42.17	20.41

We also compare our results to the best system from [2] and the relevant corresponding Gemma-2-9B-Instruct system from [27] (first two lines of Table 1). Further analysis of our cascaded systems is available in Section 4.1.

3.4. Aligned systems

When training the aligned speech encoder + connector + LLMs systems, we explore different training and inference configurations and ablating their effects. This includes the datasets used for training, connector ASR pre-training, adding LoRA to the LLM, using just the user turns in the dialogue history (for user-agent turns we insert speaker tags ‘USER:’/‘AGENT:’). The DST prompt format has the following structure for these experiments: `%speech_embeds% {"dialogue_history": %context%, "current_turn": %asr_hyp%, "domains": [%domains%], "slots": {%slots%}}`. For inference, the model completes the JSON starting with the ASR hypothesis, conditioned on the speech embeddings and the dialogue history.

We train the aligned systems with the primary goal of achieving best possible performance on SWOZ, however, we also evaluate on the MWOZ dev set to get a more complete picture of the generalization of the model. Table 1 shows the best OLMo-1B aligned DST system we obtained (A11 from Table 3). Further analysis and description of our resulting aligned systems can be found in section 4.3.

4. Results

4.1. Cascaded system analysis

Our experiments focus on ablating three main training setup factors: the training data used, the type of LLM fine-tuning, and the usage of both user and agent turns in the dialogue history. a full overview of our cascaded systems is provided in Table 2. We additionally include two text-only baseline systems from DSTC-11 [1] as reference baselines for speech-aware MWOZ.

First, we observe that fine-tuning on the original SpokenWoZ transcripts yields worse results than when using Whisper transcripts, which are of higher quality. Second, we find it beneficial to combine both SpokenWoZ and SA-MultiWOZ for training.

While full LLM fine-tuning yields the best results, we also experiment with LoRA adapters. We find that our cascaded systems achieve best results with bigger LoRA ranks, and we show the best cascade LoRA-64 system in Table 2. When scaling up the foundation LLMs or training E2E spoken DST systems, full fine-tuning becomes inconvenient. Therefore we treat the full FT cascades as the topline for the E2E LoRA systems. Lastly, while using only the user turns in the dialogue history yields good results, adding also the agent turns improves JGA on both

Table 2: *Cascaded Whisper + OLMo-1B baseline systems with two MWOZ dev reference systems from [1]*

Cascaded system configuraiton			SWOZ test		MWOZ dev	
Training data	FT	UA	JGA	SER	JGA	SER
SW-orig.	full	-	26.07	38.71	12.87	49.52
SW	full	-	28.81	34.26	12.49	52.73
MW	full	-	16.17	58.15	27.72	30.34
SW, MW	full	-	27.74	34.51	19.04	40.89
SW, MW	lora64	-	26.66	36.45	19.76	38.11
SW	full	yes	29.91	31.93	12.54	57.69
MW	full	yes	16.34	58.84	29.48	27.39
SW, MW	full	yes	30.74	31.11	23.37	34.57
ASR-DSTC large (770M params.)					25.2	
ASR-DSTC xxl (11B params.)					43.1	

datasets, resulting in our best cascaded SWOZ result of 30.74% JGA without post-processing.

4.2. Aligned system analysis

An overview of all our WavLM-large + connector + OLMo-1B aligned DST systems is shown in Table 3.

We observe that initializing the connector and speech encoder randomly and fine-tuning with LoRA (A1, A6) leads to much better results, compared to only fine-tuning the connector (A2, A7, A13). However, the two step pre-training brings additional improvements when used jointly with LoRA. Also, using both the user and agent turns in the dialogue history brings consistent improvements in JGA (e.g. A9-12).

While training only on SpokenWoZ yields reasonable results, we observe a strong tendency for overfitting on such a small dataset. We therefore find it beneficial to add SA-MultiWOZ to the training set as apart from the *profile* domain, MWOZ has the same schema as SWOZ. However, the training speech data is TTS, which is why we primarily regard this dataset as an augmentation of the DST part of the training. We obtain our best aligned OLMo-1B system by training on both datasets, and then fine-tuning on SpokenWoZ for one more epoch with a batch size of 192, achieving 34.66% JGA after post-processing (A11).

The ASR pre-training we perform focuses primarily on natural and conversational speech, and we find that our aligned models do not handle the MWOZ dev/test sets nearly as well. However, when ablating utilizing the MWOZ dev ground truth user history in the context of our model, we obtain a significant improvement in JGA (18.2 to 27.1), indicating that the utilization of a speech encoder better suited for the *human* MultiWOZ data domain would help close the performance gap between the two datasets. Similarly, when using the Whisper context for inference on SpokenWoZ instead of the previously decoded user utterances, we obtain a further increase in JGA from 31.91 (A10) to 32.89.

Lastly, we show the effect of LLM output post-processing using fuzzy matching of the slot values, which, on average, yields a further steady improvement of 3% JGA absolute.

4.3. Scaling the foundation models

Our goal with using the WavLM and OLMo models was predominantly transparency in terms of potential test data contamination. However, inspired by other works [27] we perform additional experiments with the larger Gemma-2-9B-Instruct. With the same WavLM + connector configuration and training

Table 3: Results of our aligned WavLM + connector + OLMo-1B systems. UA denotes the usage of both user and agent turns in the dialogue history, FT-sw/FT-mw denotes additional final fine-tuning on the given dataset.

#	Training data	System configuration	SpokenWoZ test		MultiWOZ dev	
			JGA[%]	SER[%]	JGA[%]	SER[%]
A1	SW	lora16 (no ASR init)	22.43	44.46	-	-
A2		connector-only	20.02	54.55	-	-
A3		lora16	27.89	33.23	-	-
A4		lora16 + UA	29.97	31.62	-	-
A5		lora16 + UA + FUZZY	32.08	29.50	-	-
A6	SW+MW	lora16 (no ASR init)	23.32	42.33	13.32	52.08
A7		connector-only	17.58	61.41	10.75	66.64
A8		lora16	27.27	35.49	16.69	46.12
A9		lora16 + UA	31.04	29.82	16.63	44.57
A10		lora16 + UA + FT-sw	31.91	28.98	-	-
A11		lora16 + UA + FT-sw + FUZZY	34.66	26.80	-	-
A12		lora16 + UA + FT-mw	-	-	17.49	43.99
A13	MW	connector-only	-	-	11.32	60.53
A14		lora16	-	-	15.34	44.96
A15		lora16 + UA	-	-	18.20	37.86
A16		lora16 + UA + FUZZY	-	-	21.62	34.29

hyperparameters as for our other models, we first run the ASR pre-training (this time with the encoder frozen as we find that unfreezing it does not yield significant improvements for larger LLMs), then we fine-tune the connector and respective LLM with LoRA $r=8$ for DST on both SWOZ and MWOZ. The results are shown in 4. Note that we also experimented with the OLMo-7B model but we omit it from these results as it did not outperform its A10/A11 OLMo-1B counterpart from Table 3.

Table 4: WavLM-large + connector + Gemma-2-9B-Instruct aligned model results on the SWOZ tes and MWOZ dev/test sets.

	JGA[%]	SER[%]
SWOZ test (FT-sw)		
Gemma-2-9B-Instruct	38.76	22.66
Gemma-2-9B-Instruct + FUZZY	42.17	20.41
MWOZ test human-verb. (FT-mw)		
Gemma-2-9B-Instruct	21.39	35.33
Gemma-2-9B-Instruct + FUZZY	24.77	32.26
MWOZ dev (FT-mw)		
Gemma-2-9B-Instruct	22.36	33.99
Gemma-2-9B-Instruct + FUZZY	25.62	30.89

The Gemma models immediately surpass SOTA on the SpokenWoZ test set with 38.76% JGA before post-processing, which yields an another 3% absolute increase to 42.17% JGA, achieving the best result among the models presented in this work. However, the FT-mw counterparts of these models do not achieve nearly as stellar performance on the MWOZ dev and human-verbatim test sets, indicating further room for improvement in terms of generalization to previously unseen named entities in the MWOZ test data.

5. Conclusions

In this work, we propose an end-to-end dialogue state tracking system based on bridging the representation spaces of a pre-

trained speech encoder with an LLM via a small transformer connector with a two-step ASR-DST fine-tuning scheme. We use open source models for both the encoder (WavLM-large) and the LLM (OLMo-1B) to mitigate the risk of test data contamination which would otherwise obstruct meaningful analysis of the scheme. The best OLMo-1B aligned model achieves 34.66% joint goal accuracy on the SpokenWoZ test set after post-processing with fuzzy matching, significantly outperforming prior work on this dataset. Finally, experiments conducted with the open-weights Gemma-2-9B-Instruct model yield the best result in this work, achieving 42.17% JGA on the SWOZ test set.

6. Acknowledgements

The work was supported by European Union’s Horizon Europe project No. SEP-210943216 "ELOQUENCE", European Defence Fund project ARCHER, Czech Ministry of Interior project No. VK01020132 "112" and by Czech Ministry of Education, Youth and Sports (MoE) through the OP JAK project "Linguistics, Artificial Intelligence and Language and Speech Technologies: from Research to Applications" (ID:CZ.02.01.01/00/23_020/0008518). Computing on IT4I supercomputer was supported by MoE through the e-INFRA CZ (ID:90254).

7. References

- [1] H. Soltau *et al.*, "DSTC-11: Speech Aware Task-Oriented Dialog Modeling Track," in *Proceedings of The Eleventh Dialog System Technology Challenge*. Prague, Czech Republic: Association for Computational Linguistics, Sep. 2023, pp. 226–234.
- [2] S. Si *et al.*, "SpokenWOZ: a large-scale speech-text benchmark for spoken task-oriented dialogue agents," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.
- [3] R. Jiang *et al.*, "Speech-Aware Multi-Domain Dialogue State Generation with ASR Error Correction Modules," in *Proceedings of The Eleventh Dialog System Technology Challenge*. Prague,

- Czech Republic: Association for Computational Linguistics, Sep. 2023, pp. 105–112.
- [4] M. Wang *et al.*, “Retrieval Augmented End-to-End Spoken Dialog Models,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 056–12 060.
- [5] Z. Chen *et al.*, “MAESTRO: Matched Speech Text Representations through Modality Matching,” in *Interspeech*, 2022, pp. 4093–4097.
- [6] W. Yu *et al.*, “Connecting Speech Encoder and Large Language Model for ASR,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 12 637–12 641.
- [7] S. Sedláček *et al.*, “Aligning Pre-trained Models for Spoken Language Translation,” 2024, arXiv:2411.18294.
- [8] Z. Chen *et al.*, “TTS4pretrain 2.0: Advancing the use of Text and Speech in ASR Pretraining with Consistency and Contrastive Losses,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7677–7681.
- [9] S. Ling *et al.*, “Adapting Large Language Model with Speech for Fully Formatted End-to-End Speech Recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 11 046–11 050.
- [10] M. Wang *et al.*, “SLM: Bridge the Thin Gap Between Speech and Text Foundation Models,” in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–8.
- [11] C. Tang *et al.*, “SALMONN: Towards Generic Hearing Abilities for Large Language Models,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [12] L. Dong *et al.*, “Integrating Speech Self-Supervised Learning Models and Large Language Models for ASR,” in *Interspeech*, 2024, pp. 3954–3958.
- [13] Z. Ma *et al.*, “An Embarrassingly Simple Approach for LLM with Strong ASR Capacity,” 2024, arXiv:2402.08846.
- [14] S. Kumar *et al.*, “Performance evaluation of SLAM-ASR: The Good, the Bad, the Ugly, and the Way Forward,” 2025, arXiv:2411.03866.
- [15] S. Chen *et al.*, “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.
- [16] D. Groeneveld *et al.*, “OLMo: Accelerating the Science of Language Models,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 15 789–15 809.
- [17] E. J. Hu *et al.*, “LoRA: Low-Rank Adaptation of Large Language Models,” in *International Conference on Learning Representations*, 2022.
- [18] P. Budzianowski *et al.*, “MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 5016–5026.
- [19] M. Eric *et al.*, “MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines,” in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 422–428.
- [20] A. Radford *et al.*, “Robust Speech Recognition via Large-Scale Weak Supervision,” in *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Jul. 2023.
- [21] T. Nekvinda and O. Dušek, “Shades of BLEU, Flavours of Success: The Case of MultiWOZ,” in *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 34–46.
- [22] J. Godfrey *et al.*, “SWITCHBOARD: telephone speech corpus for research and development,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1992.
- [23] V. Panayotov *et al.*, “Librispeech: An ASR corpus based on public domain audio books,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015.
- [24] R. Sanabria *et al.*, “How2: A Large-scale Dataset For Multimodal Language Understanding,” in *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS, 2018.
- [25] T. M. Gemma Team *et al.*, “Gemma,” 2024. [Online]. Available: <https://www.kaggle.com/m/3301>
- [26] G. Team *et al.*, “Gemma 2: Improving open language models at a practical size,” 2024. [Online]. Available: <https://arxiv.org/abs/2408.00118>
- [27] C. Richardson *et al.*, “Schema Augmentation for Zero-Shot Domain Adaptation in Dialogue State Tracking,” 2024, arXiv:2411.00150.