YUE-PUB-Speech: A Speech-based Pragmatic Understanding Benchmark for Cantonese

Yajie Wen [1,*], *Ziwei Gong* [2,*], *Chengyan Wu* [1,3,*], *Xiyun Gong* [4,**], *Yun Xue* [1], *Julia Hirschberg* [2], *Bolei Ma* [5,**]

[1] South China Normal University, [2] Columbia University, [3] Sun Yat-sen University,
[4] Guangzhou College of Commerce, [5] LMU Munich & Munich Center for Machine Learning

{yajiewen,chengyan.wu,xueyun}@m.scnu.edu.cn, {sara.ziweigong,julia}@cs.columbia.edu,
gongxiyun@gcc.edu.cn, bolei.ma@lmu.de

Abstract

Pragmatic understanding is essential for natural human communication, yet most existing studies focus on text-only settings. In this paper we introduce YUE-PUB-Speech, the first multimodal pragmatic dataset for Cantonese, a low-resource language, constructed by sampling pragmatically rich texts and recording corresponding speech by trained annotators. The corpus provides paired text–speech instances that capture pragmatic meaning in spoken language without requiring fine-grained phonological or word-level alignment. Baseline experiments show that incorporating speech signals improves pragmatic understanding compared to text-only models, highlighting the importance of multimodal resources for pragmatic research in speech and language processing. [1]

Index Terms: pragmatic understanding, computational paralinguistics, low-resource languages, under-resourced languages

1. Introduction

Pragmatic understanding—the ability to infer speaker intent beyond literal meaning—is fundamental to natural human communication [1]. In normal conversation, listeners usually rely on conversational implicature, presuppositions, reference resolution, and figurative language; these are often conveyed or disambiguated through speech-specific signals such as intonation, stress, rhythm, and timing [2]. This focus goes beyond *paralinguistics*, which studies non-lexical vocal signals, such as emotion, prosody, and voice quality, that accompany speech but are not primarily defined as meaning-bearing inferences over a discourse context [3, 4, 5, 6]. While both areas benefit from acoustic modeling, pragmatic understanding centers on *contextual meaning and inference*, where prosody can change what an utterance *implies* rather than merely how it is *delivered* [7].

Despite the centrality of spoken cues, most pragmatic benchmarks are text-only. For example, PUB [8] consolidates multiple expert-annotated datasets into a unified multiple-choice format and has become a useful testbed for pragmatic reasoning, but it does not evaluate whether models can exploit speech signals. Conversely, the speech community has developed many spoken corpora and paralinguistic benchmarks with rich audio, but they are rarely designed around pragmatic phenomena and typically lack pragmatic supervision aligned to contextual inference. This leaves a gap in evaluation: we lack benchmarks that jointly model *pragmatics* and *speech*, especially for under-resourced languages such as Cantonese, where pragmatics-oriented resources are scarce.

This gap is particularly important in low-resource settings: with limited labeled data, models have fewer opportunities to learn reliable prosody–pragmatics mappings, and pragmatic conventions are often culturally embedded, making transfer from high-resource languages unreliable. As a result, text-only evaluation can mask speech-specific failure modes, and it is unclear whether recent audio-language models achieve pragmatic competence under real low-resource conditions.

Recent multimodal and audio-language models further motivate this benchmark: combining speech and text can improve language understanding [9] yet it remains unclear when audio truly helps pragmatic inference. To address this, we introduce **YUE-PUB-Speech**, a speech-based pragmatic understanding benchmark for Cantonese. Starting from PUB [8], we translate pragmatically rich instances into Cantonese and collect matched recordings from native speakers to create aligned text–speech pairs in a unified multiple-choice question–answering form. Recordings are conducted under a consistent protocol and quality-checked for transcript–audio consistency, yielding pairs where pragmatic interpretation can draw on both lexical content and speech cues. This enables controlled comparisons between text-only, audio-only, and multimodal approaches, and supports analysis of when speech helps across pragmatic categories.

Using **YUE-PUB-Speech**, we benchmark text-only, audio-only, and text–audio models, finding that adding audio improves over text-only baselines, while audio-only models lag behind, showing that the task needs both linguistic content and speech cues; among simple fusion methods, feature concatenation with openSMILE is particularly effective for implicature. We release **YUE-PUB-Speech**, the **first** multimodal pragmatic dataset for Cantonese, with 10.87 hours of recorded speech paired with translations of PUB instances. Its unified QA formulation and evaluation suite provide a testbed for comparisons across modeling paradigms and for analyzing when speech signals improve pragmatic inference in a low-resource spoken setting.

2. Related Work

Pragmatics studies how context and speaker intent shape meaning beyond literal content (e.g., implicature, deixis) [1, 10]. Work on pragmatics has largely relied on text-only benchmarks such as PUB [8], which consolidates multiple resources into a standardized evaluation format. Yet purely textual setups can encourage "lexical dominance": models succeed by leaning on transcribed semantics while under-using the acoustic evidence carrying important meaning in speech [1, 11, 12].

[*] These authors contributed equally.
[**] indicates the corresponding author.
[1] Data available at: https://huggingface.co/datasets/Multilingual-NLP/YUE-PUB-Speech.

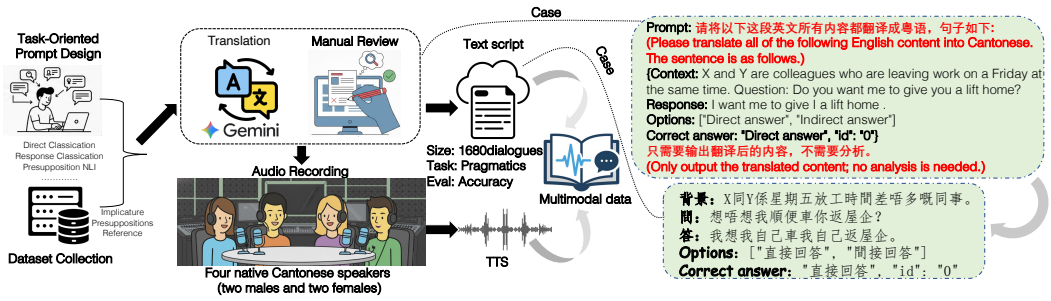

Figure 1: *The overall construction pipeline of* ***YUE-PUB-Speech*** *with an example instance.*

The speech community has established voice reasoning resources like MIAM [13] for dialogue acts and SynParaSpeech [14] for acoustic state detection. Yet, these typically focus on lower-level recognition rather than higher-order reasoning [15]. Recent benchmarks [16, 17, 12] have identified a significant voice reasoning gap, demonstrating that even models with high transcription accuracy struggle with multi-step procedural and normative reasoning in spoken interactions. This suggests that "transcribing" a signal is fundamentally distinct from "listening" to its pragmatic intent.

This scarcity in paired text–speech data for reasoning tasks is also pronounced for low-resource languages like Cantonese. Recent benchmarks such as CantoNLU [18] and HKCanto-Eval [19] began assessing Cantonese Large Language Models (LLMs) on syntax, semantics, and cultural meta-knowledge. Yet resources pairing text–speech for higher-level, pragmatics-driven reasoning are still limited [20]. We address this by introducing **YUE-PUB-Speech**, which brings PUB-style pragmatic evaluation into a spoken Cantonese setting.

3. Dataset

3.1. Dataset Construction Process

To construct YUE-PUB-Speech, we use a multimodal data collection framework (cf. Figure 1) that proceeds in 3 stages: text construction, text validation, and audio recording. We first derive and standardize pragmatically rich text instances from existing high-quality English benchmarks (Stage 1) and translate them into Cantonese using a combination of automated translation and human verification to ensure linguistic naturalness and pragmatic fidelity (Stage 2). Building upon the validated Cantonese scripts, we collect high-quality speech recordings from native speakers in a controlled acoustic environment (Stage 3). This design allows us to extend text-based pragmatic benchmarks into the spoken domain, enabling systematic study of how speech signals contribute to pragmatic understanding in Cantonese. We describe the detailed process as follows:

Stage 1: Text Data Collection. We build **YUE-PUB-Speech** upon the publicly available PUB benchmark [8], which integrates multiple expert-annotated datasets, including CIRCA [21], GRICE [22], FigQA [23], FLUTE [24], and IMPPRES [25]. Together these resources encompass a broad range of pragmatic phenomena, such as conversational implicature, presupposition, and figurative language. To ensure uniform evaluation, all data are converted into a question–answering format with answer options, enabling consistent comparison across pragmatic categories and reflecting practical discriminative scenarios in dialogue systems.

Stage 2: Text Script Construction. Our construction follows a three-stage pipeline comprising prompt construction, automated translation, and manual verification. First, to accommodate the heterogeneous formats of the original datasets (e.g., NLI and classification), we design task-specific templates that structure raw instances into a unified question–answering format, as illustrated in Figure 1. Next, we employ Gemini [26] to translate the English instances into Cantonese, covering background context, questions, and answer options while preserving pragmatic nuances. Finally, to address the low-resource nature of Cantonese and the lack of standardized norms, we conduct a dual-stage human verification process in which two native Cantonese speakers review and refine the translations to ensure linguistic naturalness and pragmatic fidelity. Instances that exhibit cultural mismatch in the Cantonese context are removed to maintain cultural validity.

This dual-stage verification process focuses on both linguistic quality and semantic consistency. In the first stage, annotators revise literal translation errors and adjust expressions to ensure idiomatic Cantonese usage, grammatical correctness, and overall naturalness. In the second stage, an additional review is performed to unify lexical choices and terminology across the dataset. When ambiguities arise, annotators consult the *Modern Cantonese Dictionary* [27] to standardize terminology and maintain consistency. Together, this mechanism ensures native-level fluency while rigorously preserving the logical structure required for pragmatic reasoning.

To assess translation quality, as shown in Table 1, we combine automatic and human evaluations. For automatic evaluation, the Cantonese translations are first back-translated into English and then compared against the original English references. We report BLEU-4 [28] and chrF++ [29] to measure lex-

Table 1: *Data quality evaluation results (text). The upper section reports automated metrics comparing model outputs against human-verified references. The lower section details human ratings averaged over 14 tasks.*

Auto Evaluation				
Task	BLEU	ChrF++	BERTScore	SBERT
Implicature	33.03	75.95	94.51	81.51
Presupposition	36.85	78.35	94.70	87.35
Reference	38.20	77.55	93.95	81.10
Human Evaluation (1–5 Scale)				
Task	Grammar	Fluency	Adequacy	Idiomatic
Implicature	3.98	4.03	3.85	3.95
Presupposition	4.10	4.10	4.10	3.97
Reference	4.04	4.10	3.92	4.04
Average	**4.01**	**4.05**	**3.90**	**3.96**

ical and character-level correspondence, and BERTScore [30] and SBERT [31] to evaluate semantic similarity. The dataset achieves consistently high scores across all metrics, with an average BERTScore above 94.0, indicating strong preservation of pragmatic meaning during translation. In addition, following established machine translation protocols [32, 33], we conduct human evaluation on the whole dataset, annotated by four native Cantonese speakers using a 5-point Likert scale across Grammar, Fluency, Adequacy, and Idiomaticity. The results show consistently high ratings (above 3.9 on average), with particularly strong performance in Idiomaticity (3.96), confirming that the translations are natural, fluent, and pragmatically faithful.

Stage 3: Audio Data Collection. After obtaining the Cantonese text, we record corresponding speech to construct aligned text–speech pairs. The process is conducted by four native Cantonese speakers (two male and two female), all university students aged 21–23, ensuring both linguistic proficiency and demographic consistency among speakers.

The dataset contains both single- and multi-speaker instances depends on task types. For single-speaker setting, contexts and corresponding questions are evenly divided between male and female speakers: each records half the background and half the questions. For multi-speaker dialogues, male and female speakers alternate, with speaker roles systematically swapped to balance speaker identity and interaction patterns.

All recordings are collected in a quiet, sound-absorbing booth (acoustic foam panels) to minimize noise. Each speaker uses a dedicated professional microphone with a pop filter for high-quality capture. For each instance, we collect a recording from native Cantonese speakers under a consistent protocol: speakers read the full dialogue context before recording; deliver a natural rendition that reflects the intended pragmatic phenomenon; and avoid exaggerated acting while still realizing pragmatically licensed prosodic contrasts. We conduct quality control and re-record samples failing these checks. When pronunciation inconsistencies arise, annotators consult standard Cantonese dictionaries to correct and unify pronunciations, ensuring consistency and linguistic accuracy across the dataset.

3.2. Dataset Statistics

The **YUE-PUB-Speech** dataset contains 1,680 Cantonese dialogue instances, comprising $\sim$11 hours of recorded speech, and spanning three core pragmatic phenomena: implicature, presupposition, and reference. The benchmark is organized into 14 tasks derived from 7 high-quality source datasets, with each task characterized by distinct dialogue structures and pragmatic demands. We group them into 3 categories:

(a) Implicature. Implicature is meaning implied rather than explicitly stated, when speakers convey information beyond the literal content of their utterance, often relying on shared context and conversational norms. Implicature-focused tasks (Tasks 1–10) form the largest dataset subset, with a total of 1,200 dialogues. These are drawn from CIRCA, GRICE, FigQA, FLUTE, and IMPPRES and are primarily designed as two-speaker conversations. Dialogue complexity varies substantially across tasks: simpler settings such as Tasks 5–7 contain short exchanges with an average of 1–2 turns per dialogue, while more complex scenarios such as Task 4 (GRICE) exhibit longer interactions, averaging 11.6 turns and 11.6 utterances per dialogue. Across all implicature tasks, dialogues contain 22.9–155.6 words on average, with dialogue durations ranging from 9.1–42.1 seconds. This diversity reflects a wide spectrum of implicature realizations, from brief conversational inferences to extended pragmatic reasoning over multi-turn discourse.

(b) Presupposition. Presupposition is information a speaker assumes true or already accepted by the listener. It remains constant even under negation and typically reflects background assumptions embedded in linguistic expressions. Presupposition understanding is covered by Tasks 11 and 12: 240 dialogues sourced from IMPPRES and DailyDialog. These tasks involve 4 distinct speakers and exhibit richer conversational structures compared to most implicature tasks. Task 11 features relatively compact dialogues with an average of 2 turns and 39.7 words per dialogue, while Task 12 contains longer conversations, averaging 5.8 turns, 7.4 utterances, and 130.4 words per dialogue. Dialogue durations range from 14.8–38.7 seconds, indicating increased contextual dependency and discourse continuity required for presupposition inference.

(c) Reference. Presupposition is information a speaker assumes is true or already accepted by the listener. It remains constant even under negation and typically reflects background assumptions embedded in linguistic expressions. Reference resolution is addressed in Tasks 13 and 14: 240 dialogues derived from GRICE and Metonymy. These involve 4 speakers and demonstrate the greatest variation in dialogue length. Task 13 has the most complex setting in the dataset, averaging 21 turns, 22 utterances, and 294.3 words per dialogue, with an average duration of 75.3 seconds. Task 14 consists of concise dialogues with only 2 turns and 42.6 words on average. This contrast captures both extended referential grounding across long conversations and localized reference phenomena in short exchanges.

Overall, **YUE-PUB-Speech** exhibits substantial diversity in dialogue length, speaker configuration, and pragmatic structure, making it a challenging and realistic benchmark for evaluating pragmatic understanding in spoken Cantonese.

4. Experiments and Results

To evaluate the usability of the **YUE-PUB-Speech** dataset for pragmatic analysis, we investigate 3 representative categories of baseline methods: single-modality models using either text or speech features alone; an LLM-based framework concatenating audio encoder representations with textual inputs; and audio language models (ALMs) evaluated under a zero-shot setting. This design enables a systematic comparison across different modeling paradigms and allows us to examine the contribution of speech signals to pragmatic understanding.

4.1. Setup

To evaluate the proposed dataset, we select a set of models covering open-source and closed-source ALMs and LLMs, and commonly used speech encoders. The selected open-source models include *Qwen2.5-7B-Instruct* [34], *SALMONN* [35], *Qwen2.5-Omni* [36], and *WavLLM* [37], together with the closed-source *Gemini-2.5-Pro* [38]. These models represent different design paradigms for language and multimodal reasoning. For speech representation, we consider a range of audio encoders with varying modeling characteristics: handcrafted acoustic features (eGeMAPSv02) extracted by *openSMILE* [39], self-supervised speech encoders such as *Whisper-small* [40], *wav2vec* [41], and *HuBERT-based* [42] models, and a *HuBERT-Chinese* model with strong acoustic representation learning. By combining different LLMs and audio encoders, we aim to systematically analyze the impact of model architecture and speech representations on pragmatic understanding. We report accuracy as the primary evaluation metric for all tasks.

Table 2: ***YUE-PUB-Speech*** *Dataset General Statistics. Benchmark covers 3 major pragmatic categories from 7 high-quality source datasets. Size shows number of dialogue instances in each task; source shows original benchmark from which the task is derived.*

Task ID	Size	Hours	Phenomena	Source	#Speakers	#Labels	#dialogs	#words	#utterances	#turns	Avg utt/dialog	Avg words/dialog	Avg dialog dur(sec)	Task Type
Task 1-2	240	1	Implicature	CIRCA	2	[2,5]	240	10348	360	240	3	43.13	14.94	Response Classication
Task 3	120	0.67	Implicature	CIRCA	2	5	120	7946	480	360	4	66.22	20.24	Response Classication
Task 4	120	1.4	Implicature	GRICE	2	4	120	18666	1392	1392	11.6	155.55	42.09	Implicature Recovery
Task 5-6	240	0.76	Implicature	FigQA	2	2	240	7784	240	240	2	32.43	11.46	Agreement / Sarcasm Detection
Task 7	120	0.3	Implicature	FLUTE	2	[2,3]	120	2752	120	120	1	22.93	9.14	Figurative Language Understanding
Task 8-9	240	1.64	Implicature	FLUTE	2	[2,3]	240	19008	240	240	2	79.2	24.66	Figurative Language Understanding
Task 10	120	0.41	Implicature	IMPPRES	2	3	120	3635	240	240	2	30.29	12.29	Implicature NLI
Task 11	120	0.49	Presupposition	IMPPRES	4	3	120	4760	240	240	2	39.67	14.84	Presupposition NLI
Task 12	120	1.29	Presupposition	DailyDialog	4	2	120	15643	891	706	7.425	130.36	38.70	Presupposition over QA
Task 13	120	2.51	Reference	GRICE	4	2	120	35311	2640	2520	22	294.26	75.29	Deicitc QA
Task 14	120	0.4	Reference	Metonymy	4	4	120	5108	240	240	2	42.57	12.12	Reference via Metonymy
Total	1,680	10.87	-	-	-	-	1,680	130,961	7,383	6,538	-	-	-	-

4.2. Experiments

We conduct experiments under 3 settings to evaluate pragmatic understanding on **YUE-PUB-Speech**, to systematically compare different modeling paradigms and analyze the impact of speech signals on pragmatic understanding:

Unimodal inference: We investigate unimodal inference under both text-only and speech-only conditions. For text-only, we consider 2 configurations based on the Qwen2.5-7B-Instruct model: zero-shot prompting and supervised fine-tuning (SFT) with LoRA, where the SFT results are reported using 5-fold cross-validation. For the speech modality, we evaluate zero-shot inference using Qwen2.5-Omni, in which only raw speech is provided as input. This aims to assess the individual contribution of text and speech modalities.

Text–audio feature concatenation: We adopt a text–audio feature concatenation framework under an SFT paradigm, extracting acoustic representations with different speech encoders; concatenate them with the text input; and fine-tune Qwen2.5-7B-Instruct via LoRA-based SFT. Performance is evaluated via 5-fold cross-validation. This isolates the effect of different speech encoders under the same LLM and supervision.

End-to-end ALMs: We evaluate open- and closed-source ALMs that take both speech and text input and perform pragmatic reasoning in a zero-shot regime. This benchmarks current multimodal capabilities without task-specific training.

All experiments were run on 2 $\times$ NVIDIA A6000 GPUs with 48GB memory each. For generation, we used a sampling configuration with a 1.0 top-p value and a 0.2 temperature. For supervised fine-tuning, we adopted LoRA with a rank of 8, alpha value of 16, and dropout rate of 0.05 across all SFT settings. All SFT experiments are evaluated using 5-fold cross-validation, with a 4:1 train/test split within each fold.

4.3. Results and Discussion

Table 3 presents the main experimental results on YUE-PUB-Speech. Across all architectures, Presupposition consistently emerges as the most challenging category, yielding lower scores than Implicature and Reference due to its reliance on deep logical deduction of unstated assumptions. In terms of baselines, the SFT method proves critical for pragmatic alignment. Text_{sft} significantly outperforms zero-shot approaches (gaining nearly 10% in Implicature), demonstrating that SFT is essential for capturing specific conversational Cantonese maxims raw embeddings miss. For reasoning-intensive tasks (Presupposition and Reference), Gemini-2.5-Pro's dominance in these metrics (peaking at 82.50% on Reference) underscores the unique advantage of end-to-end multimodal architectures, which effectively synergize advanced logical reasoning with cross-modal contextualization to solve complex pragmatic inferences. However, in the Implicature tasks, the lightweight text-audio feature concatenation with openSMILE achieves the highest accuracy (80.33%), outperforming even larger ALMs. This shows the effectiveness of **YUE-PUB-Speech** in low-resource settings: incorporating audio consistently boosts performance. Crucially, it indicates that **YUE-PUB-Speech** dataset high-quality training data enables efficient, specialized models to surpass generalist Audio-Language Models in detecting subtle prosodic cues.

Table 3: *Main experimental results on YUE-PUB-Speech, comparing unimodal, text–audio concatenation, and ALM approaches for pragmatic understanding. Imp., Pre., and Ref. correspond to the 3 pragmatic categories: Implicature, Presupposition, and Reference, respectively. Avg. denotes the overall accuracy averaged across all tasks.*

Setting	Imp.	Pre.	Ref.	Avg.
Unimodal Baselines				
$\text{Audio}_{zero-shot}$	50.83	37.50	58.75	50.06
$\text{Text}_{zero-shot}$	55.64	52.42	60.50	56.19
Text_{sft}	65.79	56.42	61.50	61.24
Text–Audio Feature Concatenation (Qwen2.5-7B as LLM)				
Whisper-small	63.17	54.58	66.25	62.38
hubert-based	68.50	58.75	68.33	67.08
wav2vec	71.33	62.50	69.58	69.82
hubert-chinese	71.50	62.92	70.00	70.06
openSMILE	80.33	66.67	75.00	77.62
Audio-Language Models				
WavLLM	53.00	48.33	50.00	51.46
SALMONN	60.48	53.75	47.92	56.97
Qwen2.5-omni	70.58	59.17	67.92	68.57
Gemini-2.5-Pro	74.33	73.75	82.50	75.42

5. Conclusions and Limitations

We introduce **YUE-PUB-Speech**, the first multimodal pragmatic dataset for Cantonese that aligns pragmatically rich text with recorded speech to support pragmatic understanding in spoken language. By focusing on a low-resource language, the dataset addresses a critical gap in existing pragmatic research, which has largely been limited to text-only resources. Experimental results demonstrate that incorporating speech signals consistently improves pragmatic understanding over text-only baselines, underscoring the importance of multimodal modeling for Cantonese pragmatics. Nevertheless, our current evaluation relies on relatively simple baseline settings and does not yet explore more advanced multimodal integration strategies, which may limit the full potential of the dataset. *In future work*, we plan to expand the dataset to cover a broader range of pragmatic phenomena, speakers, and speech styles, and to further explore advanced multimodal approaches for pragmatic reasoning.

6. Generative AI Use Disclosure

The authors acknowledge the use of ChatGPT for language editing (grammar and clarity) .

7. Acknowledgements

This work is in part supported by the funds provided NSF DBI-2229929 (The NSF AI Institute for Artificial and Natural Intelligence). YW, CW, and YX are supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011370, the National Natural Science Foundation of China (32371114), the Characteristic Innovation Projects of Guangdong Colleges and Universities (No. 2018KTSCX049), and the Guangdong Provincial Key Laboratory (No. 2023B1212060076).

8. References

[1] B. Ma, Y. Li, W. Zhou, Z. Gong, Y. J. Liu, K. Jasinskaja, A. Friedrich, J. Hirschberg, F. Kreuter, and B. Plank, "Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 8679–8696. [Online]. Available: https://aclanthology.org/2025.acl-long.425/

[2] N. G. Ward, *Prosodic Patterns in English Conversation*. Cambridge: Cambridge University Press, 2019.

[3] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. MacIntyre, and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks," in *Interspeech 2020*, 2020, pp. 2042–2046.

[4] H. Kim, S. Seo, K. Jeong, O. Kwon, S. Kim, J. Kim, J. Lee, E. Song, M. Oh, J.-W. Ha, S. Yoon, and K. M. Yoo, "Paralinguistics-aware speech-empowered large language models for natural conversation," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 131 072–131 103. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/ecfa5340fd896e5314bc5e132b5dd5ca-Paper-Conference.pdf

[5] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Interspeech 2010*, 2010, pp. 2794–2797.

[6] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *Interspeech 2013*, 2013, pp. 148–152.

[7] J. Hirschberg, "Communication and prosody: Functional aspects of prosody," *Speech Communication*, vol. 36, no. 1, pp. 31–43, 2002, eSCA Workshop on Dialogue and Prosody, September 1999.

[8] S. Sravanthi, M. Doshi, P. Tankala, R. Murthy, R. Dabre, and P. Bhattacharyya, "PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities," in *Findings of the Association for Computational Linguistics: ACL 2024*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 12 075–12 097. [Online]. Available: https://aclanthology.org/2024.findings-acl.719/

[9] B. Wang, X. Zou, G. Lin, S. Sun, Z. Liu, W. Zhang, Z. Liu, A. Aw, and N. F. Chen, "AudioBench: A universal benchmark for audio large language models," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, L. Chiruzzo, A. Ritter, and L. Wang, Eds. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 4297–4316. [Online]. Available: https://aclanthology.org/2025.naacl-long.218/

[10] G. Kasper, "Speech acts in interaction: Towards discursive pragmatics," *Pragmatics and language learning*, vol. 11, 2006.

[11] S. Floyd, O. Jouravlev, M. Poliak, Z. Mineroff, E. Gibson, and E. Fedorenko, "Three distinct components of pragmatic language use: Social conventions, intonation, and world knowledge–based causal reasoning," *Proceedings of the National Academy of Sciences*, vol. 122, no. 50, p. e2424400122, 2025. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2424400122

[12] W. Yang, Y. Li, Y. Wei, M. Fang, and L. Chen, "Speechr: A benchmark for speech reasoning in large audio-language models," 2025. [Online]. Available: https://arxiv.org/abs/2508.02018

[13] P. Colombo, E. Chapuis, M. Labeau, and C. Clavel, "Code-switched inspired losses for spoken dialog representations," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 8320–8337. [Online]. Available: https://aclanthology.org/2021.emnlp-main.656

[14] Z. Zhang, W. Xu, Z. Dong, K. Wang, Y. Wu, J. Peng, R. Wang, and D.-Y. Huang, "Paralbench: A large-scale benchmark for computational paralinguistics over acoustic foundation models," *IEEE Transactions on Affective Computing*, vol. 16, no. 3, pp. 1290–1306, 2025.

[15] K. Anuranjana, S. Mallepally, S. Mareddy, A. Shukla, and R. Mamidi, "Survey on computational approaches to implicature," in *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, S. Lalitha Devi and K. Arora, Eds. AU-KBC Research Centre, Chennai, India: NLP Association of India (NLPAI), Dec. 2024, pp. 224–229. [Online]. Available: https://aclanthology.org/2024.icon-1.25/

[16] Y. Lin, Z. Hu, Q. Wang, Y. Liu, H. Zhang, J. Subramanian, N. Vlassis, H. H. Li, and Y. Chen, "Voice evaluation of reasoning ability: Diagnosing the modality-induced performance gap," 2025. [Online]. Available: https://arxiv.org/abs/2509.26542

[17] X. Zhifei, M. Lin, Z. Liu, P. Wu, S. Yan, and C. Miao, "Audio-reasoner: Improving reasoning capability in large audio language models," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 23 829–23 851. [Online]. Available: https://aclanthology.org/2025.emnlp-main.1216/

[18] J. Min, Y. H. Ng, S. Chan, H. S. Zhao, and E.-S. A. Lee, "CANTONLU: A benchmark for cantonese natural language understanding," 2025. [Online]. Available: https://arxiv.org/abs/2510.20670

[19] T. C. Cheng, C. S. Cheng, C.-m. Lau, E. Lam, W. C. Yat, H. O. Yu, and C. H. Chong, "HKCanto-eval: A benchmark for evaluating Cantonese language understanding and cultural comprehension in LLMs," in *Proceedings of the 29th Conference on Computational Natural Language Learning*, G. Boleda and M. Roth, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 1–11. [Online]. Available: https://aclanthology.org/2025.conll-1.1/

[20] P. Xie and K. Chen, "Developing a multilingual dataset and evaluation metrics for code-switching: A focus on hong kong's polylingual dynamics," 2025. [Online]. Available: https://arxiv.org/abs/2310.17953

[21] A. Louis, D. Roth, and F. Radlinski, ""I'd rather just go to bed": Understanding indirect answers," in *Proceedings of the 2020 Conference on Empirical Methods in Natural*

Language Processing (EMNLP), B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 7411–7425. [Online]. Available: https://aclanthology.org/2020.emnlp-main.601/

[22] Z. Zheng, S. Qiu, L. Fan, Y. Zhu, and S.-C. Zhu, "GRICE: A grammar-based dataset for recovering implicature and conversational rEasoning," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 2074–2085. [Online]. Available: https://aclanthology.org/2021.findings-acl.182/

[23] E. Liu, C. Cui, K. Zheng, and G. Neubig, "Testing the ability of language models to interpret figurative language," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 4437–4452. [Online]. Available: https://aclanthology.org/2022.naacl-main.330/

[24] T. Chakrabarty, A. Saakyan, D. Ghosh, and S. Muresan, "FLUTE: Figurative language understanding through textual explanations," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 7139–7159. [Online]. Available: https://aclanthology.org/2022.emnlp-main.481/

[25] P. Jeretic, A. Warstadt, S. Bhooshan, and A. Williams, "Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 8690–8705. [Online]. Available: https://aclanthology.org/2020.acl-main.768/

[26] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, D. Silver, M. Johnson, I. Antonoglou, J. Schrittwieser, A. Glaese, J. Chen, E. Pitler, T. Lillicrap, A. Lazaridou *et al.*, "Gemini: A family of highly capable multimodal models," 2025. [Online]. Available: https://arxiv.org/abs/2312.11805

[27] G. Huang, A. Gorin, J. Gauvain, and L. Lamel, "Machine translation based data augmentation for cantonese keyword spotting," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*. IEEE, 2016, pp. 6020–6024. [Online]. Available: https://doi.org/10.1109/ICASSP.2016.7472833

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040/

[29] M. Popović, "chrF++: words helping character n-grams," in *Proceedings of the Second Conference on Machine Translation*, O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, and J. Kreutzer, Eds. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 612–618. [Online]. Available: https://aclanthology.org/W17-4770/

[30] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with BERT," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[31] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. [Online]. Available: https://aclanthology.org/D19-1410/

[32] Y. Chen, Z. Song, X. Wu, D. Wang, J. Xu, J. Chen, H. Zhou, and L. Li, "MTG: A benchmark suite for multilingual text generation," in *Findings of the Association for Computational Linguistics: NAACL 2022*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 2508–2527. [Online]. Available: https://aclanthology.org/2022.findings-naacl.192/

[33] C. Wu, B. Ma, Y. Liu, Z. Zhang, N. Deng, Y. Li, B. Chen, Y. Zhang, Y. Xue, and B. Plank, "M-ABSA: A multilingual dataset for aspect-based sentiment analysis," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 2530–2557. [Online]. Available: https://aclanthology.org/2025.emnlp-main.128/

[34] Q. Team, "Qwen2.5: A party of foundation models," September 2024. [Online]. Available: https://qwenlm.github.io/blog/qwen2.5/

[35] T. Changli, Y. Wenyi, S. Guangzhi, C. Xianzhao, T. Tian, L. Wei, L. Lu, M. Zejun, and Z. Chao, "SALMONN: Towards generic hearing abilities for large language models," *arXiv:2310.13289*, 2023.

[36] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, B. Zhang, X. Wang, Y. Chu, and J. Lin, "Qwen2.5-omni technical report," 2025. [Online]. Available: https://arxiv.org/abs/2503.20215

[37] S. Hu, L. Zhou, S. Liu, S. Chen, L. Meng, H. Hao, J. Pan, X. Liu, J. Li, S. Sivasankaran, L. Liu, and F. Wei, "Wavllm: Towards robust and adaptive speech large language model," in *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, ser. Findings of ACL, Y. Al-Onaizan, M. Bansal, and Y. Chen, Eds., vol. EMNLP 2024. Association for Computational Linguistics, 2024, pp. 4552–4572. [Online]. Available: https://doi.org/10.18653/v1/2024.findings-emnlp.263

[38] G. Team, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *CoRR*, vol. abs/2507.06261, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2507.06261

[39] F. Eyben, M. Wöllmer, and B. W. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, A. D. Bimbo, S. Chang, and A. W. M. Smeulders, Eds. ACM, 2010, pp. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[40] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[41] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 3465–3469. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-1873

[42] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3122291