Words at Play: Benchmarking Audio Pun Understanding in Large Audio-Language Models

Anonymous ACL submission

Abstract

Puns represent a typical linguistic phenomenon that exploits polysemy and phonetic ambiguity to generate humour, posing unique challenges for natural language understanding. Within pun research, audio plays a central role in human communication except text and images, while datasets and systematic resources for spoken puns remain scarce, leaving this crucial modality largely underexplored. In this paper, we present APUN-Bench, the first benchmark dedicated to evaluating large audio language models (LALMs) on audio pun understanding. Our benchmark contains 4,434 audio samples annotated across three stages: pun recognition, pun word location and pun meaning inference. We conduct a deep analysis of APUN-Bench by systematically evaluating 9 state-of-the-art LALMs, uncovering substantial performance gaps in recognizing, localizing, and interpreting audio puns. This analysis reveals key challenges, such as positional biases in audio pun location and error cases in meaning inference, offering actionable insights for advancing humour-aware audio intelligence.

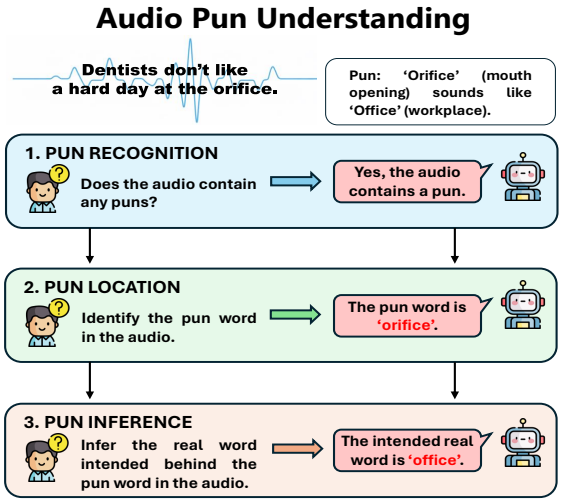

Figure 1: An example of the pun understanding in recognition, location and inference.

1 Introduction

As a typical linguistic phenomenon, puns exploit the polysemy of language and create cognitive tension through the complex interplay of phonetics, orthography, and generating humour (Partington, 2009). It is not only a vehicle for humorous expression, but also embodies a vital manifestation of human creativity and communicative flexibility (Aleksandrova, 2022), with its inherent ambiguity makes it a rigorous test of true language understanding. Miller et al. (2017) and Sun et al. (2022a) show that understanding pun phenomenon usually requires crossing three levels: determining whether there is a pun in the language fragment, locating the specific pun, and inferring the implied alternative words or double meanings. In terms of applications, puns play a significant role in areas such as education, voice assistants, and entertainment, thus they also provide a critical benchmark for evaluating the capabilities of general artificial intelligence (Morris et al., 2024).

Audio puns play a particularly important role in interpersonal communication (Karpenko, 2017; Palmann and Miller, 2025). Unlike text or images, audio is the most common medium of human communication, which carries not only linguistic content but also adds layers of depth and flexibility through acoustic cues. Moreover, audio puns are often shaped by phonetic ambiguity, pronunciation similarity, and acoustic variability (Benzeghiba et al., 2007), endowing them with rich expressive value in communication. Systematic investigation of audio puns therefore contributes to a deeper understanding of human language and cognitive mechanisms, while simultaneously presenting new challenges and opportunities for the development of audio understanding and multimodal artificial intelligence.

Against this backdrop, benchmarking audio puns is critical for revealing the limitations of current LALMs and for improving their ability to process this complex linguistic phenomenon, while related research remains markedly underexplored. In recent years, researchers in natural language processing and computer vision have introduced several pun benchmarks (Ouyang et al., 2024; Xu et al., 2024; Zhang et al., 2024) to systematically evaluate the recognition, explanation, and generation capabilities of large language models (LLMs) in text and image pun tasks, advancing the understanding of written and visual puns. However, systematic studies and publicly available datasets addressing puns at the phonetic level are still lacking. Meanwhile, existing state-of-the-art LALMs have focused primarily on relatively basic evaluation tasks, such as answering speech questions and multimodal understanding (speech–video) (Yang et al., 2024; Sakshi et al., 2024; Wang et al., 2024; Gao et al., 2024; Zhang et al., 2025; Wang et al., 2025). While these tasks are essential for testing fundamental audio comprehension, they do not specifically address more complex linguistic phenomena such as puns. This evaluation gap constrains the potential of LALMs to advance toward higher-level audio understanding.

In this work, we introduce APUN-Bench, a benchmark for audio pun understanding that comprises 4,434 audio samples, with complete annotations spanning three stages: pun recognition, pun word location, and pun meaning inference. The first two stages are identification tasks, requiring models to determine whether a given audio clip contains a pun and to further locate the punning words. Unlike previous research that has primarily focused on binary classification or word identification, our design additionally evaluates models' ability to perform coarse-grained location in audio, thereby revealing their strengths and weaknesses at different levels. The third stage, pun meaning inference, assesses whether models can correctly infer the secondary meanings of three types of puns: heterographic, homographic, and homophonic. Additionally, we propose the pun dataset which integrates two parts: synthetic data and real-world speech data. The synthetic component is created by converting existing textual pun corpora into speech samples, while the real-world component is collected from publicly available speech resources to capture authentic speech contexts. We review and evaluate the quality of the speech data and corresponding annotations to ensure the overall reliability of the dataset. Our evaluation results on APUN-Bench reveal that current LALMs still have significant gaps in the recognition and understanding of audio puns. Overall, our key contributions are as follows:

- We propose APUN-Bench, the benchmark for audio pun understanding, including 4,434-sample dataset combining synthetic and real-world speech with semi-automatic annotation and human verification.
- We comprehensively cover different stages of pun understanding, including pun recognition, pun word location, and pun meaning inference, as well as evaluate systematically on 9 open-source and proprietary LALMs using APUN-Bench.
- We provide an in-depth analysis of model performance, revealing key insights such as models' positional bias in pun location and common error types in pun inference, which point to potential directions for future improvements.

2 Related Work

2.1 Benchmarks for LALMs

A growing number of benchmarks have been developed to evaluate LALMs. General-purpose resources such as Audiobench Wang et al. (2024) and Airbench Yang et al. (2024) focus on recognition and generative comprehension across diverse acoustic conditions. More recent work extends this scope: Sakshi et al. (2024) and Wang et al. (2025) assess multi-step spoken understanding and reasoning, while Zhang et al. (2025) highlights challenges in natural conversational settings. Other studies investigate specific dimensions, including cross-modal hallucination (Sung-Bin et al., 2024), dialogue ambiguity (Gao et al., 2024), and reasoning strategies for structured inference (Ma et al., 2025; Xie et al., 2025).

2.2 Pun Studies

Puns, which rely on polysemy, homonymy and semantic conflict, are one of the most challenging linguistic phenomena for NLP models. Existing research on pun understanding has largely concentrated on recognition (Miller et al., 2017; Diao et al., 2018; Zhou et al., 2020; Jentzsch and

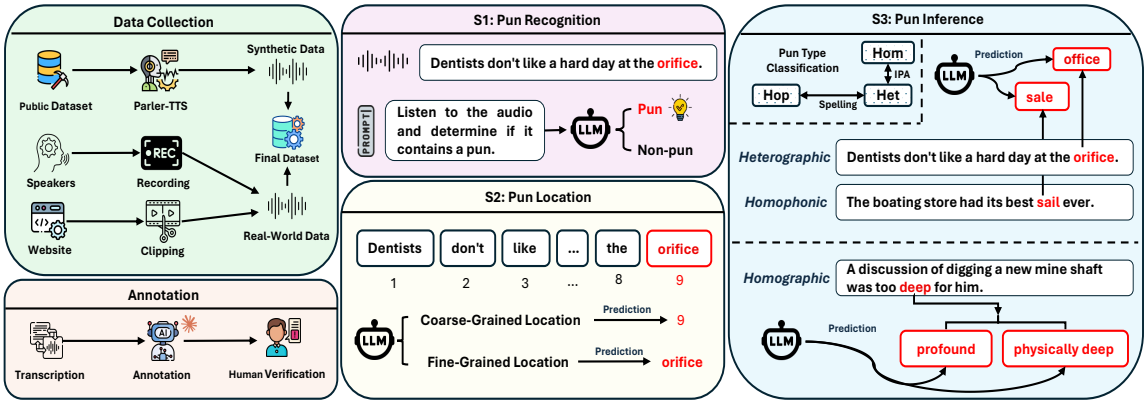

Figure 2: The Overview of APUN-Benchmark, covering data construction, human annotation and three evaluation stages: (S1) Pun Recognition, detecting the presence of an audio pun; (S2) Pun Location, identifying the specific pun word in the audio; and (S3) Pun Inference, inferring its dual meanings.

Kersting, 2023), location (Zou and Lu, 2019), and explanation (Sun et al., 2022a), but has also begun to explore multimodal extensions such as visual puns (Zhang et al., 2024). Beyond understanding, scholars have further investigated automatic pun generation (Yu et al., 2018; Luo et al., 2019; Sun et al., 2022b; Su et al., 2025) and systematic evaluation of the ability of large-scale language models to understand puns (Xu et al., 2024). However, as a central form of human communication, speech remains largely unexplored in pun research, and there is an urgent need for relevant benchmarks and systematic studies.

3 Benchmark

Our goal is to evaluate the abilities of LALMs in understanding puns. Specifically, we aim to analyze the detection and deep semantic reasoning when processing puns in spoken language of current LALMs and provide insights for building more robust models with enhanced language understanding. As existing benchmarks focus primarily on pun research within textual modalities, we introduce the APUN-Bench, designed to measure a model's ability to detect and infer pun-related phenomena in speech. Figure 2 provides an overview of the benchmark, including data collection, human annotation and task descriptions.

Overview of APUN-Bench. Inspired by pun research on unimodal text, APUN-Bench comprises three main stage categories: pun recognition, pun word location, and pun meaning inference. Pun recognition is a binary classification task to determine whether a speech segment contains a pun, aligning with prior work (Xu et al., 2024). The pun location stage focuses on the specific pun word, including localization and identification subtasks. The pun meaning inference stage evaluates understanding of the secondary meaning, with subtasks designed for homophonic, heterographic, and homographic puns (Su et al., 2025). In this study, compound puns that involve multiple punning words within the same sentence are not considered. In total, APUN-Bench consists of 4,434 audio clips, including 2,971 audio puns paired with their corresponding pun words and inferred alternative words, as well as 1,463 negative samples.

3.1 Data Collection

The dataset construction consists of two main components: (1) Synthetic Audio Dataset: Audio samples synthesized from existing pun text data; (2) Real-world Audio Dataset: Audio data collected from real-world sources, including human recordings and publicly available websites, followed by editing and annotation.

Synthetic Audio Dataset. One challenge in constructing the benchmark is collecting audio data and performing fine-grained annotations. To address this, we used the existing SemEval 2017 pun text dataset (Miller et al., 2017), which comprises 2,389 positive instances (puns) and 1,152 negative instances (non-puns). The positive set comprises instances annotated with both the punning word and its corresponding replacement. Then, we employ state-of-the-art speech synthesis Parler-TTS (Lacombe et al., 2024; Lyth and King, 2024) to randomly assign each text to one of 34 distinct synthetic speakers, ensuring diversity in speaker

Dataset	Heterographic	Homophonic	Homographic	Negative	Overall
Synthetic Dataset	794	304	1298	1,152	3,541
- Segment Clip (s)	4.18	4.36	3.88	3.47	3.97
- Sentence Length	12.21	13.89	11.68	8.50	11.57
Real-World Dataset	223	72	287	311	582
- Segment Clip (s)	6.36	6.16	5.34	6.73	6.14
- Sentence Length	11.12	11.40	11.55	10.88	11.23

Table 1: APUN-Bench data statistics, including average audio clip duration and sentence length.

characteristics. More details can be found at Appendix C.

Real-world Audio Dataset. In addition to the synthetic audio dataset, we collect publicly available videos and manually segment the speech content from the *O. Henry Museum Pun-Off* [1], a long-standing global pun competition where contestants must create puns within a limited time to compete for the championship. To address the limited availability of public audio data, we further collect data from pun websites [2] and create realistic pun speech by manual recording. Some details can be found in Appendix E.1. Furthermore, we constructed negative samples from real speech. The dataset contains 582 positive samples and 311 negative samples. These annotated data effectively introduce the challenge of understanding puns in audio, requiring models to accurately detect and interpret them under complex conditions. In total, our dataset comprises 3,541 synthetic and 892 real-world audio samples, ensuring diversity across data sources. Table 1 presents the statistics of both synthetic and real-world data.

3.2 Dataset Construction

We provide detailed descriptions for 3 different stages, comprising a total of 6 subtasks.

Stage 1: Pun Recognition. This task assesses the model's ability to detect the presence of puns in audio, which serves as a prerequisite for more fine-grained comprehension and reasoning tasks. Inspired by (Xu et al., 2024), we construct the question format that include both pun and non-pun to mitigate potential biases, whose primary question format is "Determine whether the given audio is a pun/non-pun".

Stage 2: Pun Location. This task evaluates the ability of LALMs to localize pun words within spoken sentences. We investigate whether a model that can already detect the presence of a pun in audio may still struggle to accurately identify the specific pun word. To enable more granular evaluation, we divide this task into two levels: (1). Coarse Location: The model is evaluated for its ability to accurately locate the approximate location of pun words in the audio, without requiring the semantic accuracy of the identified words. Even if the model locates a semantically inaccurate word, as long as its location makes sense, it is considered correct. (2). Fine-Grained Location: this subtask further requires the model to accurately identify actual puns word in the audio.

For real-world speech data, manually labeling each audio segment in fine-grained detail is time-consuming. Previous research has demonstrated that leveraging LLMs to assist in dataset construction is both efficient and feasible (Gong et al., 2023; Liu et al., 2023; Hyun et al., 2024). Xu et al. (2024)) has demonstrated that Claude ranks among the best existing models for understanding text-based puns. Accordingly, we first transcribe the audio into text using the Whisper-large model (Radford et al., 2022), and then employ Claude-Opus (particularly Claude-Opus-4-1) (Anthropic, 2025) to identify and locate the puns. To enhance its task comprehension and annotation accuracy, we prompt Claude-Opus with a small set of manually annotated examples along with feedback. More details can be found in the appendix D.

Stage 3: Pun Inference. This task evaluates the ability of LALMs to infer the intended meaning of puns in spoken sentences. Based on the characteristics, puns are classified into three categories for evaluation: heterographic, homophonic, and homographic: (1) Heterographic puns involve words with similar but not identical pronunciations. For this type, the model needs to infer the correct pronunciation and the corresponding intended meaning substitute based on the given pun; (2) Homophonic puns involve words with identical pronunciations but different spellings. It is similar to heterographic puns, where the task is to identify the intended alternative word. (3) Homographic puns involve words with the same written form but different meanings. The model is required to generate phrases that capture both meanings and compare them with the gold standard, thereby assessing its ability to understand polysemy.

Based on the characteristics of different pun categories, we adopt a differentiated classification strategy. For homographic puns, which rely on lexical polysemy without spelling differences, we classify them by checking the consistency of word spellings. For homophonic and heterographic puns, whose

[1] https://www.punoffatx.brushsquaremuseums.org/
[2] https://pun.me/

defining feature lies in the phonetic differences between the punning word and its substitute, we transcribe the target words into International Phonetic Alphabet (IPA) symbols and determine their category based on whether their pronunciations match.

To infer replacement words for heterographic and homophonic puns, we employed a similar process to Stage 2: given a pun, Claude-Opus is prompted to infer and generate suitable replacements. It is worth noting that for homographic puns, which involve polysemy, we provide Claude-Opus with targeted prompts to generate two distinct sets of phrases representing different meanings of the pun. All outputs are subsequently verified and manually corrected to mitigate potential model bias and produce the final gold-standard annotations.

Human Annotation. During benchmark construction, we conduct multiple rounds of manual development and reviews. In the raw data collection phase, each audio clip is rigorously examined each audio clip for pun presence, pun words, categories, and replacement words to ensure the quality of the audio pun data. Audio samples with substandard quality are re-collected to ensure data reliability. In addition, when inferring double meanings for homographic puns in the synthetic speech dataset, we manually review the generated polysemous outputs, revise inappropriate content, and correct errors to obtain gold-standard annotations. The manual error correction rate is approximately 6%, and all checks are performed by the authors. Furthermore, a final human verification process is conducted to ensure the overall reliability of the dataset. More details can be found at Appendix E.2

4 Evaluation

Baseline Models. We evaluate APUN-Bench on a range of 9 recent LALMs, widely cited in the literature and capable of jointly processing text and audio information, including Qwen2-Audio-Instruct (Chu et al., 2024), Audio-Reasoner(Xie et al., 2025), Qwen-2.5-omni (Jin Xu, 2025), MiniCPM (Yao et al., 2024), Gemini 2.0 Flash (Jaech et al., 2024), SALMONN (Tang et al., 2023), GPT4o-Audio (Hurst et al., 2024), Omni-R1 (Rouditchenko et al., 2025) and MERaLiON2 (MERaLiON Team, 2024; Huang et al., 2025). Each model is configured with a temperature of 0 to ensure deterministic outputs.

Evaluation Strategy. We adopt appropriate evaluation metrics for different tasks. For pun recognition tasks, we use accuracy, precision, recall, and F1 score as evaluation criteria, following the common practice in previous pun recognition research (Gepalova et al., 2024; Xu et al., 2024). Inspired by (Sung-Bin et al., 2024), we also report the proportion of "pun" responses produced by each model to examine potential biases in pun recognition.

For the pun word location stage, the coarse-grained setting requires only identifying the approximate location of the pun within the speech. Specifically, we first transcribe the target audio and obtain the IPA representation of each word, then compute the character-level edit distance between these representations and the IPA of the pun word predicted by the model. In addition, we incorporate phonetic characteristics by defining sets of vowels and consonants, mapping phonemes into feature vectors, and calculating their cosine similarity. We then combine these two similarity measures to select the word position with the highest overall similarity as the prediction result, as shown in the formula below:

$$S_{\text{edit}}(w_i) = 1 - \frac{d_{\text{edit}}(\phi(w_i), \phi(\hat{w}))}{\max(|\phi(w_i)|, |\phi(\hat{w})|)}, \quad (1)$$

$$S_{\cos}(w_i) = \cos\big(v(\phi(w_i)), v(\phi(\hat{w}))\big), \quad (2)$$

$$S(w_i) = S_{\text{edit}}(w_i) + S_{\cos}(w_i), \quad (3)$$

where $\phi(\cdot)$ denotes the IPA representation of a word, d_{edit} is the Levenshtein distance, and $v(\cdot)$ maps phonemes into feature vectors. For fine-grained location, an exact match is required to identify puns within speech. We first perform lemmatization on the words and then directly match them with the pun words predicted by the model to obtain the final evaluation outcome.

During the pun inference phase, results for heterographic and homophonic puns are evaluated by exact matching between the model-predicted pun words and the ground-truth annotations, as illustrated in the formula below:

$$\text{Eval}(p) = \begin{cases} 1, & \hat{w} = w^*, \\ 0, & \text{otherwise}, \end{cases} \quad (4)$$

where $\hat{w}$ is the model-predicted pun word and w^* is the ground-truth alternative words. For homographic puns, as described in Section 3.2, we first construct a gold standard capturing the dual meanings of each pun. The model's predicted meanings are then compared with this gold standard using

similarity matching, and the two pairs with the highest similarity are selected. The formula is illustrated as below:

$$\text{Eval}(p) = \begin{cases} 1, & S(\hat{m}_{j_1}, m^*_{i_1}) > \tau \ \wedge \ S(\hat{m}_{j_2}, m^*_{i_2}) > \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

where m^*_i are the gold-standard meanings of a homographic pun, $\hat{m}_j$ are the model-predicted meanings, $S(\cdot, \cdot)$ denotes the similarity function, and τ is the threshold. Based on empirical experiments, pairs with similarity scores above 0.3 (τ = 0.3) are retained and regarded as positive examples for this task.

5 Results and discussion

Based on our proposed APUN-Bench, we conduct a systematic evaluation of LALMs in pun understanding. We highlight several key findings from the experimental results, analyze the potential underlying causes of the observed phenomena and explore promising directions for improvement.

5.1 Main Results

We systematically evaluate baseline models across the three stages of APUN-Bench on both synthetic and real-world datasets, with the results summarized in Table 2. In the pun recognition stage, most baseline models perform above chance level (50% accuracy). However, with the exception of GPT4o-Audio, Gemini2.0-Flash, and MERaLiON2, their recall rates are generally low, suggesting that the models tend to be overly conservative in identifying puns and are prone to false negatives. In the pun location stage, models are able to identify the approximate positions of more than half of the puns in the coarse-grained location task. Nevertheless, their performance drops substantially in the fine-grained detection task when recovering the exact pun words, revealing LALMs' limited ability to accurately transcribe specific pun words. In the pun inference stage, models achieve markedly better results on homophonic puns than on heterographic puns, consistent with findings that homophony constitutes only about 2–3% of the English lexicon (Marian et al., 2012), whereas a four-letter English word has on average about 10.33 phonological neighbors.

Moreover, proprietary models consistently outperformed open-source speech models in APUN-Bench, especially in pun location and inference, underscoring the gap between the two in deep semantic understanding. We further conduct McNemar's test to examine whether the results are statistically significant. For example, GPT4o-Audio achieves a highly significant advantage over MiniCPM in fine-grained pun location ($p < 1 \times 10^{-10}$). Overall, although closed-source models exhibit certain advantages across tasks, current auditory language models still face significant challenges in detecting and understanding puns, with overall performance remaining unsatisfactory.

5.2 Are LALMs Biased Toward Certain Types of Pun Word Location?

In the pun location stage, different models show notable performance gaps between coarse and fine-grained locations, with differences of around 20% being common across models. While this indicates that the models are generally consistent in transcribing and recognizing audio puns, we further explore their recognition capabilities across specific pun categories. Table 3 presents the performance of various models on three categories of puns, evaluated through both coarse-grained and fine-grained tasks in synthetic dataset. The results reveal that most models perform significantly better on homographic puns, excelling in both location detection and exact matching. For heterographic and homophonic puns, although their accuracy in coarse location is comparable, the accuracy of fine-grained location for homophonic puns is substantially lower. This finding highlights that existing models still encounter considerable difficulties in accurately reproducing fully homophonic puns.

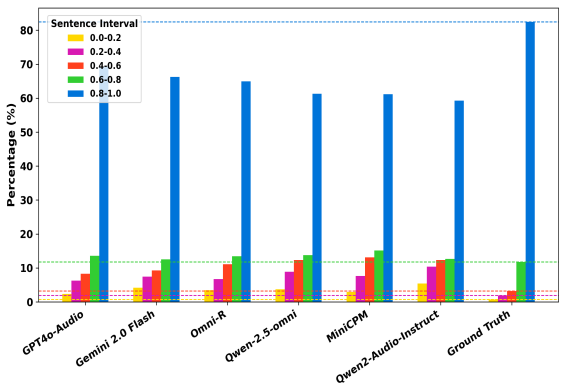

Figure 3: Statistical comparison of predicted in different LALMs and ground-truth pun positions in APUN-Bench, showing the distribution of pun words across different sentence intervals (beginning, middle, end of the sentence).

Model	Size	Pun Recognition					Pun Location		Pun Inference		
		Acc. (↑)	Pre (↑)	Rec (↑)	F1 (↑)	Pun (%)	Coa (↑)	Fin (↑)	Heg (↑)	Hog (↑)	Hop (↑)
Synthetic Dataset											
Qwen2-Audio-Instruct	7B	45.78	71.70	39.45	50.90	38.06	50.81	31.77	13.47	17.91	26.97
SALMONN	13B	51.91	72.51	52.60	60.97	51.81	34.41	22.35	15.23	24.63	25.32
Audio-Reasoner	7B	55.27	72.62	59.81	65.59	54.47	50.39	32.31	23.05	27.25	37.50
Qwen-2.5-omni	7B	64.04	78.71	68.07	73.00	61.76	60.98	48.01	40.55	32.38	60.19
MiniCPM	8.7B	62.31	78.00	65.77	71.37	27.92	59.94	44.62	39.54	37.36	67.43
Omni-R1	8.9B	65.28	77.39	72.58	74.91	66.97	63.49	49.43	44.20	39.73	63.48
MERaLiON2	10B	73.40	75.66	92.56	83.26	**87.37**	60.44	43.78	50.88	29.43	79.93
Gemini 2.0 Flash ✯	-	**80.00**	**91.04**	79.86	85.08	62.39	65.80	48.34	52.39	53.29	**92.11**
GPT4o-Audio ✯	-	78.04	78.48	**95.92**	**86.33**	83.87	**75.97**	**56.63**	**59.06**	**57.81**	85.19
Real-world Dataset											
Qwen2-Audio-Instruct	7B	58.64	95.11	37.61	53.90	25.53	49.05	24.95	23.87	21.59	26.02
SALMONN	13B	44.86	57.07	60.41	58.69	68.56	28.82	14.11	18.38	23.15	16.43
Audio-Reasoner	7B	70.99	89.97	61.94	73.37	41.14	48.71	25.98	34.97	25.70	38.35
Qwen-2.5-omni	7B	71.68	**99.09**	56.79	72.21	37.12	59.21	42.85	50.00	35.78	57.53
MiniCPM	8.7B	74.35	98.08	61.62	75.68	40.80	54.73	38.03	43.69	34.86	54.79
Omni-R1	8.9B	83.50	97.37	76.59	85.74	50.95	58.86	41.31	44.59	35.00	57.53
MERaLiON2	10B	**88.18**	87.05	96.04	**91.32**	**71.46**	54.56	35.45	49.54	32.28	62.50
Gemini 2.0 Flash ✯	-	87.51	84.82	**98.51**	91.15	60.98	**67.52**	**51.03**	**62.78**	47.36	**82.85**
GPT4o-Audio ✯	-	85.36	87.02	92.56	89.71	66.89	66.95	50.08	60.98	**50.70**	76.71

Table 2: Evaluation Results of the Synthetic and Real-World Datasets on APUN-Bench. We evaluate various LALMs on our proposed AVHBench. ✯ represents the proprietary LALMs.

Model	Heterographic		Homophonic		Homographic	
	Coa (↑)	Fin (↑)	Coa (↑)	Fin (↑)	Coa (↑)	Fin (↑)
SALMONN	32.37	17.12	35.86	8.55	35.32	28.81
Qwen2-Audio-Instruct	52.14	29.47	44.41	13.81	51.51	37.41
Audio-Reasoner	50.50	29.34	39.47	11.18	52.90	39.11
Qwen-2.5-omni	53.53	36.39	56.91	21.38	66.54	61.42
MiniCPM	56.30	34.13	62.50	20.06	61.58	56.85
Omni-R1	58.06	38.16	60.53	21.71	67.54	62.89
MERaLiON2	55.16	31.36	62.50	17.10	63.21	57.70
Gemini 2.0 Flash ✯	60.96	36.90	70.72	35.19	67.62	58.48
GPT4o-Audio ✯	**74.06**	**44.20**	**76.64**	**42.43**	**76.99**	**67.62**

Table 3: Evaluation results of the Synthetic Dataset on APUN-Bench. We evaluate various LALMs on our proposed APUN-Bench.

5.3 Analyzing the Position of Pun Words in Sentences

According to the incongruity theory of humour, punchlines typically appear at the end of a sentence (Shahaf et al., 2015). Building on this, we analyze the positional preferences of different models in identifying puns and compared them with manual annotation results, as shown in Figure 3. Sentence Interval refers to the normalized sentence position; for example, 0–0.2 denotes the first 20% of the whole sentence. The annotation data indicate that the vast majority of puns occur within the last 40% of a sentence, with fewer than 10% located in the first 60%, further supporting the incongruity theory. In contrast, most model predictions exhibit a more uniform distribution: although the last 20% still contains the highest proportion of predicted puns, more than 20% appear in the first 60%. These findings suggest that guiding models to focus more on the latter half of a sentence may improve their performance in pun location.

To examine this hypothesis, we conduct a simple experiment. When designing prompts for pun location, we introduce the information that strengthens the model's attention to the latter half of the sentence. The following shows an improved segment of the prompt:

Improved Segment of Prompt

[...] Identify the pun word in the audio, paying particular attention to the latter half of the sentence. [...]

We conduct evaluations on Qwen-2.5-omni-7B and MiniCPM, achieving 63.49% and 62.32% accuracy in coarse-grained pun location, respectively, surpassing the baseline models by 2.51% and 2.38% percentage points, as shown in Table 2. For fine-grained location, the models achieve 49.56% and 46.42% accuracy, outperforming the baselines by 1.55% and 1.80% percentage points, respectively. While the overall improvements are limited, the results indicate that emphasizing the latter part of a sentence may contribute to better pun localization and provide initial insights for fu-

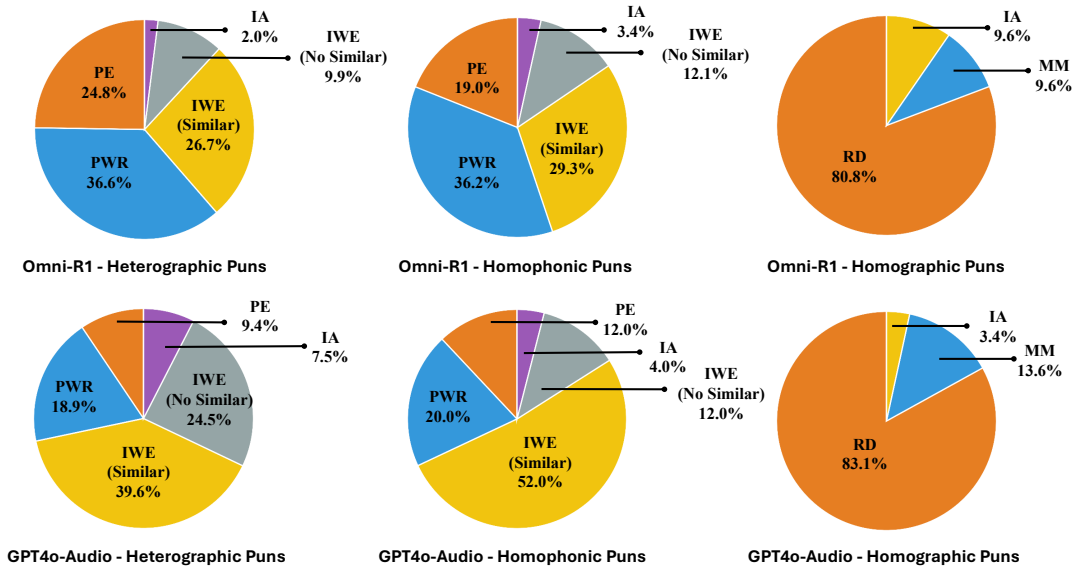

Figure 4: The distribution of error types in pun inference for Omni-R1 and GPT4o-Audio across heterographic, homographic, and homophonic puns.

ture research.

5.4 Audio Pun Inference: Where do LALMs Fall Short?

Figure 4 presents the distribution of error types for Omni-R1 and GPT4o-Audio across 450 instances, with 150 examples for each pun category. Detailed definitions of error types are provided in Table 4. For heterographic and homophonic puns, the dominant error type in both models is Word Inference Error (including Similar and No Similar), accounting for more than one-third of the errors. Similarly, for homographic puns, the primary error type shifts to Reasoning Disorder, which constitutes over three-quarters of the errors, while Meaning Mismatch accounts for only about 10%. This indicates that although the models are generally capable of perceiving speech and identifying one layer of pun meaning, they encounter substantial difficulties in uncovering and restoring the secondary meaning. Moreover, for heterographic and homophonic puns, GPT4o-Audio exhibits a markedly lower proportion of Pun Word Repeat errors compared to Omni-R1, suggesting that larger-scale models are less prone to rigidly reproducing the pun word itself as output. Finally, Irrelevant Answer represents the smallest proportion of errors across all categories and both models, implying that most models are able to understand the task requirements and attempt to provide answers related to pun reasoning rather than completely unrelated outputs. Overall, our error analysis suggests that enhancing speech comprehension is crucial for improving performance. Building on prior research, this can be achieved through the targeted design of audio chain-of-thought (Ma et al., 2025) reasoning mechanisms and by expanding training data to strengthen speech–text modal alignment (Liu et al., 2024).

6 Conclusion

In this paper, we introduce APUN-Bench, an audio pun evaluation benchmark designed to comprehensively evaluate pun understanding in LALMs. We provide the first audio pun dataset, automate evaluation metrics and conduct extensive experiments on 9 open-source and proprietary models. Furthermore, by covering three stages: pun recognition, pun word location, and pun meaning inference, our benchmark systematically evaluates the model's capabilities across multiple dimensions of audio pun understanding. The experiment results show that while some models achieve promising performance on basic recognition tasks, they continue to face substantial challenges in pun word fine-grained location and inference, particularly for heterographic and homophonic puns. Through detailed analyses of positional biases and error types, we highlight the unique obstacles that audio puns pose and point to future directions for advancing more robust and human-like audio language understanding.

Limitations

While this research offers new perspectives and insights into the understanding of audio puns, several limitations remain: (1) The scope of pun types extends beyond the three categories examined in this study, with more complex forms such as recursive and compound puns remaining outside its consideration. (2) Puns often arise in more complex speech environments, such as multi-turn dialogues, whereas our dataset is restricted to single-sentence instances. (3) The size of the real-world corpus still remains limited, which constrains the statistical robustness of our findings. However, despite these limitations, this paper represents the first systematic study of verbal puns, and its data and analyses offer valuable references for subsequent research on the intersection of language and speech understanding.

References

Elena Aleksandrova. 2022. Pun-based jokes and linguistic creativity. *The European Journal of Humour Research*, 10(1):88–107.

Anthropic. 2025. System card: Claude opus 4 & claude sonnet 4. https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf.

Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al. 2007. Automatic speech recognition and speech variability: A review. *Speech communication*, 49(10-11):763–786.

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

Yufeng Diao, Hongfei Lin, Di Wu, Liang Yang, Kan Xu, Zhihao Yang, Jian Wang, Shaowu Zhang, Bo Xu, and Dongyu Zhang. 2018. Weca: A wordnet-encoded collocation-attention network for homographic pun recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2507–2516.

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. 2024. Benchmarking open-ended audio dialogue understanding for large audio-language models. *arXiv preprint arXiv:2412.05167*.

Arina Gepalova, Adrian-Gabriel Chifu, and Sébastien Fournier. 2024. Clef 2024 joker task 1: exploring pun detection using the t5 transformer model. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024). CEUR Workshop Proceedings*, pages 1857–1861.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2023. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.

Xin Huang, Tarun Kumar Vangani, Minh Duc Pham, Xunlong Zou, Bin Wang, Zhengyuan Liu, and Ai Ti Aw. 2025. Meralion-textllm: Cross-lingual understanding of large language models in chinese, indonesian, malay, and singlish. *Preprint*, arXiv:2501.08335.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Lee Hyun, Kim Sung-Bin, Seungju Han, Youngjae Yu, and Tae-Hyun Oh. 2024. SMILE: Multimodal dataset for understanding laughter in video with language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Sophie Jentzsch and Kristian Kersting. 2023. Chatgpt is fun, but it is not funny! humor is still challenging large language models. *arXiv preprint arXiv:2306.04563*.

Jinzheng He Hangrui Hu Ting He Shuai Bai Keqin Chen Jialin Wang Yang Fan Kai Dang Bin Zhang Xiong Wang Yunfei Chu Junyang Lin Jin Xu, Zhifang Guo. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Hanna Karpenko. 2017. Humor as a part of interpersonal communication. *Journal of Vasyl Stefanyk Precarpathian National University. Series of social and human sciences*, (4, no. 1):195–199.

Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi. 2024. Parler-tts. https://github.com/huggingface/parler-tts.

Zhenyu Li, Minghao Zhao, Tian Xu, Rui Chen, Xinyu Yang, Hao Wu, and Shanshan Wang. 2024. Meralion-audiollm: Bridging audio and language with large language models. *arXiv preprint arXiv:2412.09818*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Fuli Luo, Shunyao Li, Pengcheng Yang, Baobao Chang, Zhifang Sui, Xu Sun, et al. 2019. Pun-gan: Generative adversarial network for pun generation. *arXiv preprint arXiv:1910.10950*.

Dan Lyth and Simon King. 2024. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *Preprint*, arXiv:2402.01912.

Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. 2025. Audio-cot: Exploring chain-of-thought reasoning in large audio language model. *arXiv preprint arXiv:2501.07246*.

Viorica Marian, James Bartolotti, Sarah Chabal, and Anthony Shook. 2012. Clearpond: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities.

MERaLiON Team. 2024. Meralion-audiollm: Bridging audio and language with large language models. *Preprint*, arXiv:2412.09818.

Tristan Miller, Christian F Hempelmann, and Iryna Gurevych. 2017. Semeval-2017 task 7: Detection and interpretation of english puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68.

Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 2024. Position: Levels of agi for operationalizing progress on the path to agi. In *Forty-first International Conference on Machine Learning*.

Kun Ouyang, Yuanxin Liu, Shicheng Li, Yi Liu, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2024. Punchbench: Benchmarking mllms in multimodal punchline comprehension. *arXiv preprint arXiv:2412.11906*.

Anna Palmann and Tristan Miller. 2025. What's in a pun? assessing the relationship between phonological distance and perceived funniness of punning jokes. *HUMOR*, (0).

Alan Scott Partington. 2009. A linguistic account of wordplay: The lexical grammar of punning. *Journal of Pragmatics*, 41(9):1794–1809.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. 2025. Omni-r1: Do you really need audio to fine-tune your audio llm? *arXiv preprint arXiv:2505.09439*.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1065–1074.

Yuchen Su, Yonghua Zhu, Ruofan Wang, Zijian Huang, Diana Benavides-Prado, and Michael Witbrock. 2025. A survey of pun generation: Datasets, evaluations and methodologies. *arXiv preprint arXiv:2507.04793*.

Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Tagyoung Chung, Jing Huang, Yang Liu, and Nanyun Peng. 2022a. Expunations: Augmenting puns with keywords and explanations. *arXiv preprint arXiv:2210.13513*.

Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Shuyang Gao, Tagyoung Chung, Jing Huang, Yang Liu, and Nanyun Peng. 2022b. Context-situated pun generation. *arXiv preprint arXiv:2210.13522*.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. 2024. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. *arXiv preprint arXiv:2410.18325*.

Yunxin Tang, Yusheng Liu, Zhiyong Wu, Jiajun Li, Qian Wang, Zhongkai Meng, Zheng Zhang, Lei Xie, and Ming Wang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2309.07985*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-reasoner: Improving reasoning capability in large audio language models. *Preprint*, arXiv:2503.02318.

Zhijun Xu, Siyu Yuan, Lingjie Chen, and Deqing Yang. 2024. " a good pun is its own reword": Can large language models understand puns? *arXiv preprint arXiv:2404.13599*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v:

A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. A neural approach to pun generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1660.

Jian Zhang, Linhao Zhang, Bokai Lei, Chuhan Wu, Wei Jia, and Xiao Zhou. 2025. Wildspeech-bench: Benchmarking audio llms in natural speech conversation. *arXiv preprint arXiv:2506.21875*.

Tuo Zhang, Tiantian Feng, Yibin Ni, Mengqin Cao, Ruying Liu, Katharine Butler, Yanjun Weng, Mi Zhang, Shrikanth S Narayanan, and Salman Avestimehr. 2024. Creating a lens of chinese culture: A multimodal dataset for chinese pun rebus art understanding. *arXiv preprint arXiv:2406.10318*.

Yichao Zhou, Jyun-Yu Jiang, Jieyu Zhao, Kai-Wei Chang, and Wei Wang. 2020. " the boating store had its best sail ever": Pronunciation-attentive contextualized pun recognition. *arXiv preprint arXiv:2004.14457*.

Yanyan Zou and Wei Lu. 2019. Joint detection and location of english puns. *arXiv preprint arXiv:1909.00175*.

A Baseline Models

Qwen2-Audio-Instruct. Chu et al. (2024) present Qwen2-Audio, a large audio–language model that integrates an audio encoder with the Qwen LLM to process spoken queries, dialogues, and instruction-following tasks. The instruct variant evaluated here follows the official release, focusing on robust audio understanding across speech and environmental sound domains.

SALMONN. Tang et al. (2023) introduce SALMONN, which equips LLMs with general "hearing" abilities. It combines speech, audio event, and music encoders with a frozen LLM backbone, supporting tasks ranging from ASR and speech translation to audio captioning. SALMONN has been widely adopted as an open-source baseline for broad auditory competence.

Audio-Reasoner. Xie et al. (2025) propose Audio-Reasoner, a reasoning-enhanced model trained with structured chain-of-thought supervision and a large corpus of annotated audio–text pairs (CoTA). It is designed to strengthen multi-step deduction and temporal reasoning, offering explicit reasoning traces alongside competitive performance.

Qwen-2.5-Omni. Jin Xu (2025) extend the Qwen family with an omni-modal series capable of processing text, audio, and vision within a unified interface. The 3B and 7B variants are optimised for instruction following, open-domain spoken QA, and multimodal reasoning, providing strong baselines at different parameter scales.

MiniCPM. Yao et al. (2024) present MiniCPM-V, a lightweight multimodal model optimised for efficient deployment on mobile devices. Despite its compact scale, MiniCPM demonstrates performance comparable to GPT-4V on several multimodal tasks. Its audio-capable versions support real-time speech understanding for instruction-following scenarios.

Omni-R1. Rouditchenko et al. (2025) explore whether direct audio fine-tuning is necessary for competitive performance. Omni-R1 leverages large-scale instruction tuning on text paired with synthetic audio, achieving strong spoken understanding and reasoning without explicit raw-audio fine-tuning.

MERaLiON-AudioLLM. Li et al. (2024) introduce MERaLiON-AudioLLM, a multilingual and multimodal large audio–language model. The family emphasises robustness to multilingual inputs, accents, and code-switching. In our evaluation, we reference the 10B variant (*MERaLiON2*), which extends this line with improved large-scale instruction-following capabilities.

GPT-4o-Audio. Hurst et al. (2024) document the GPT-4o family; we evaluate the audio-enabled version that accepts spoken input for transcription, dialogue, and multimodal reasoning. As a closed-source model, GPT-4o-Audio sets a strong commercial baseline for real-time, latency-sensitive applications.

Gemini 2.0 Flash. The Gemini Flash models are proprietary systems released by Google DeepMind as part of the Gemini family Jaech et al. (2024). They are optimised for low-latency instruction following with speech input. We include Gemini 2.0 as a strong closed-source baseline to contextualise open-source performance.

B Error Types

We summarize the error types of LALMs in the audio pun inference stage, as presented in Table 4.

Position Error (PE). The model fails to provide the correct word corresponding to the given pun word and may instead return a word from another position in the sentence.

Pun Word Repeat (PWR). The model outputs the given pun word itself rather than its intended

meaning.

Word Inference Error with Similar (WIE). The model infer an incorrect word, often manifested as an other word with a similar pronunciation.

Word Inference Error with No Similar (WIE). The model infer an incorrect word, often manifested as another word with a completely different pronunciation.

Irrelevant Answer (IA). The model can only infer a single layer of meaning, where one pair of matches shows high similarity while the other pair fails to align.

Reasoning Disorder (RD). The model generates double meaning words or phrases that deviate from the intended meanings defined in the gold standard.

Meaning Mismatch (MM). The model's answer is irrelevant style of inference meaning or incorrect styles for the pun word.

C Synthesis Process

We use Parler-TTS for speech synthesis. This open-source, lightweight text-to-speech system is capable of generating high-quality, natural-sounding speech conditioned on various speaker characteristics, such as gender, pitch, and speaking style (Lacombe et al., 2024). To ensure diversity, we utilize the 34 default speaker embeddings and randomly assign one to each generated utterance, as shown in Table 5. Finally, a manual filtering step is conducted to screen the missing words or unclear pronunciations to improve the overall quality of the speech data.

D Annotation Prompts

We use the following prompt to instruct Claude-Opus (Claude-Opus-4-1) to identify and locate puns in transcribed speech. The prompt includes a clear task description, output specification, and a few manually annotated examples with corrective feedback to improve the accuracy of the auxiliary annotation. Some details are shown as Figure 3.

E Human Involvement

E.1 Annotation

For real-world data recordings, we recruit 10 adult anonymous participants with strong English-speaking backgrounds, all of whom hold a college education and represent diverse majors from different countries. To ensure diversity in speech, each participant records approximately 30–60 pun sentences, and we provide a compensation of $0.3 per sentence recording.

E.2 Human Verification

Human verification is carried out by two experts, each with a master's or doctoral degree and strong English proficiency. The review process is divided into two parts, including both synthetic dataset and real-world dataset:

Audio Quality: Experts examine the audio files to identify any missing or under-recorded segments and to ensure that the recordings are clear and intelligible. Each audio clip averages no more than 8 seconds in length. To preserve the authenticity of natural speech, accent variation is not restricted. For quality assessment, we randomly sample 300 audio files from the dataset. Approximately 1.5% of the samples require re-recording due to defects, yielding a low overall defect rate and confirming the high quality of the audio data.

Annotation Verification: After the initial annotation stage, we conduct a systematic quality check. Experts confirm the presence of puns, specify alternatives for heterographic and homophonic puns, and validate the accuracy of dual meanings in homographic puns. A random sample of 300 data points is selected for cross-validation. On average, reviewers spend 45–60 seconds evaluating each speech sample. Each entry is independently verified by each experts, and inter-annotator agreement is measured using *Cohen's K*, which reaches 0.92, indicating substantial consistency across annotators.

E.3 Human Evaluation

We randomly sample 60 data points for manual evaluation from each stage to compare the performance of current mainstream LALMs with that of human participants. Task-specific questionnaires are designed and distributed to three anonymous volunteers via the scientific research platform Prolific [3], targeting primarily native English speakers from the UK and the USA. Each participant receives a reward of $5.4 per completed questionnaire. Table 6 presents the evaluation results, indicating that while some state-of-the-art LALMs achieve performance comparable to human evaluators in certain pun understanding tasks, they still lag significantly behind humans in others, such as pun

[3] https://www.prolific.com/

Error Type	Definition	Example (Audio Transcription)	Prediction	Ground Truth
Heterographic Puns & Homophonic Puns				
Position Error (PE)	The model fails to provide the correct word corresponding to the given pun word and may instead return a word from another position in the sentence.	When asked to picture the perfect modern defensive weapon the Claymore springs to **mine**.	Claymore	mind
Pun Word Repeat (PWR)	The model outputs the given pun word itself rather than its intended meaning.	Did you hear about the nervous preacher ? He had sweaty **psalms**.	psalms	palm
Word Inference Error (WIE) (Similar)	The model infer an incorrect word, often manifested as another word with a similar pronunciation.	One ear of corn said to the other'You're getting **husky**'.	hoarse	husk
Word Inference Error (WIE) (No Similar)	The model infer an incorrect word, often manifested as another word with a completely different pronunciation.	George Westinghouse was a refrigerator **magnate**.	entrepreneur	magnet
Irrelevant Answer (IA)	The model's answer is unrelated to the pun inference for the pun word.	She was only a Butcher's daughter, but there wasn't much more she could **loin**.	line human given an audio ...	learn
Homographic Puns				
Reasoning Disorder (RD)	The model can only infer a single layer of meaning, where one pair of matches shows high similarity while the other pair fails to align.	How could I trust the ceiling fan installer when I knew he was always **screwing up**.	making mistakes AND causing trouble	making mistakes AND installing screws
Meaning Mismatch (MM)	The model generates double-meaning words or phrases that deviate from the intended meanings defined in the gold standard.	Can honeybee abuse lead to a **sting** operation?	steamboat AND steamship	undercover police activity AND painful insect bite
Irrelevant Answer (IA)	The model's answer is irrelevant style of inference meaning or incorrect styles for the pun word.	Then there was the occasion I spotted a health - oriented cafe displaying a sign reading' California **Shakes**.' Obviously, I thought.	The two meanings of 'shake' in this context	earthquake AND milkshake

Table 4: Error types of pun inference in heterographic, homophonic and homographic puns

Laura	Gary	Jon	Lea
Karen	Rick	Brenda	David
Eileen	Jordan	Mike	Yann
Joy	James	Eric	Lauren
Rose	Will	Jason	Aaron
Naomie	Alisa	Patrick	Jerry
Tina	Jenna	Bill	Tom
Carol	Barbara	Rebecca	Anna
Bruce	Emily		

Table 5: List of the 34 speakers used in Parler-TTS.

inference. In addition, we observe that human evaluators exhibit a conservative bias when identifying puns, tending to classify utterances as non-puns. However, it is worth noting that pun understanding relies on rich linguistic and world knowledge, representing a high-level language comprehension task. Therefore, this task calls for larger-scale and more diverse human evaluations, which we plan to pursue in future work.

E.4 Permission

The Institutional Review Board (IRB) of our institution has approved the human studies presented in this paper. In addition, we have obtained permission to use data collection from the public pun website *O. Henry Museum Pun-Off*.

/* Definition */
You are an expert linguist with specialized knowledge of pun research. Your role is to annotate transcribed spoken sentences by identifying the pun word. A pun word is a word or phrase that exploits multiple meanings, sound similarities, or semantic ambiguities to create humor or wordplay.

/* Instruction */
For each given sentence, identify the pun word. If there is no pun word, output "None". Only output the pun word (or the minimal pun phrase), without explanations or additional text. Be consistent with the format shown in the examples.

/* Examples */
Pun Sentence: Dentists don't like a hard day at the orifice.
Output: orifice
Pun Sentence: A discussion of digging a new mine shaft was too deep for him.
Output: deep
...

/* Test Data */
Pun Sentence: He didn't tell his mother that he ate some glue. His lips were sealed.
Output:

Figure 5: Prompts used to guide Claude-Opus-4-1 in pun location for auxiliary annotation.

/* Definition */
You are an expert linguist with specialized knowledge of pun research. Your role is to annotate transcribed spoken sentences by inferring the alternative word that the pun word replaces or plays upon. A pun word is a word or phrase that exploits multiple meanings, sound similarities, or semantic ambiguities to create humor or wordplay.

/* Instruction */
For each given pun sentence, infer the intended alternative word. Only output the alternative word (or minimal phrase) without explanations or additional text. Ensure consistency with the format shown in the examples.

/* Examples */
Pun Sentence: Dentists don't like a hard day at the orifice.
Pun word: orifice
Output: office
...

/* Test Data */
Pun Sentence: Geologists can be sedimental about their work.
Output: sedimental
Output:

Figure 6: Prompts used to guide Claude-Opus-4-1 in pun inference for auxiliary annotation (for heterographic and homophonic puns).

Model	Pun Recognition		Pun Location		Pun Inference		
	Acc. (↑)	Pun (%)	Coa (↑)	Fin (↑)	Heg (↑)	Hog (↑)	Hop (↑)
SALMONN	63.33	50.00	43.33	23.33	18.33	28.33	18.33
Qwen2-Audio-Instruct	51.66	43.33	58.33	33.33	30.00	23.63	28.33
Audio-Reasoner	58.33	50.00	63.33	40.00	31.67	34.61	35.42
Qwen-2.5-omni	78.33	58.33	60.00	36.67	43.33	31.66	65.00
MiniCPM	53.33	33.33	53.33	30.00	46.67	44.06	58.33
Omni-R1	75.00	65.00	56.67	33.33	45.00	44.82	63.89
MERaLiON2	73.33	**86.67**	56.67	26.67	51.67	36.66	65.00
Gemini 2.0 Flash ★	**81.67**	63.33	70.00	41.66	51.67	53.33	81.67
GPT4o-Audio ★	73.33	86.67	73.33	46.67	53.33	52.54	78.33
Human	76.67	41.67	**76.67**	**60.00**	**78.33**	**73.33**	**90.00**

Table 6: Comparison with Human Evaluation and LALMs

/* Definition */
You are an expert linguist with specialized knowledge of pun research. Your role is to annotate transcribed spoken sentences by inferring the two distinct meanings of a homographic pun word. A pun word is a word or phrase that exploits multiple meanings, sound similarities, or semantic ambiguities.

/* Instruction */
For each given pun sentence, identify the pun word and infer its two distinct meanings. Output exactly two replacement words or phrases, each representing one meaning of the pun word. Provide only the two items, separated by a comma, with no additional text. Ensure consistency with the format shown in the examples.

/* Examples */
Pun Sentence: A discussion of digging a new mine shaft was too deep for him
Pun Word: deep
Output: profound, physically deep
...

/* Test Data */
Pun Sentence: If at first you don't succeed then skydiving is not for you.
Pun Word: succeed
Output:

Figure 7: Prompts used to guide Claude-Opus-4-1 in pun inference for auxiliary annotation (for homographic puns).