KoALa-Bench: Evaluating Large Audio Language Models on Korean Speech Understanding and Faithfulness

Anonymous ACL submission

Abstract

Recent advances in large audio language models (LALMs) have enabled multilingual speech understanding. However, benchmarks for evaluating LALMs remain scarce for non-English languages, with Korean being one such underexplored case. In this paper, we introduce **KoALa-Bench**, a comprehensive benchmark for evaluating Korean speech understanding and speech faithfulness of LALMs. In particular, KoALa-Bench comprises six tasks. Four tasks evaluate fundamental speech understanding capabilities, including automatic speech recognition, speech translation, speech question answering, and speech instruction following, while the remaining two tasks evaluate speech faithfulness, motivated by our observation that several LALMs often fail to fully leverage the speech modality. Furthermore, to reflect Korea-specific knowledge, our benchmark incorporates listening questions from the Korean college scholastic ability test as well as content covering Korean cultural domains. We conduct extensive experiments across six models, including both white-box and black-box ones. Our benchmark, evaluation code, and leaderboard are publicly available at https://ksbench.github.io/Korean-Benchmark/.

1 Introduction

Recent advances in multimodal large language models (MLLMs) (Chu et al., 2024; Tang et al., 2024; Rubenstein et al., 2023; Team et al., 2023) have enabled models to process diverse modalities, including images, videos, and audio. Among these, speech is particularly important as it represents one of the most natural forms of human interaction. Accordingly, large audio language models (LALMs) (Team, 2026; Xu et al., 2025; Fang et al., 2025; Chu et al., 2024), which integrate speech encoders with pretrained language models, have emerged as a promising paradigm for speech understanding and spoken interaction, increasingly supporting multilingual speech inputs.

As LALMs expand their multilingual capabilities, comprehensive evaluation frameworks for speech understanding across diverse languages become increasingly important. However, existing benchmarks remain predominantly English-centric, leaving languages such as Korean largely underexplored (Wang et al., 2025; Gao et al., 2025; Yang et al., 2024). Meanwhile, existing Korean speech benchmarks are primarily designed for traditional speech processing tasks rather than evaluating the speech understanding capabilities of LALMs (Bang et al., 2020; Ha et al., 2020; Zeroth Project, 2018). As a result, there is currently a lack of standardized benchmarks specifically designed to evaluate LALMs using Korean speech inputs. To address this limitation, we introduce **Ko**rean **A**udio **La**nguage benchmark, called **KoALa-Bench**, a comprehensive benchmark for evaluating the Korean speech understanding capabilities of LALMs.

KoALa-Bench consists of six tasks, four for standard speech understanding and two newly proposed for evaluating speech faithfulness. The four standard tasks cover automatic speech recognition (ASR), speech translation (ST), speech question answering (SQA), and speech instruction following (SIF). For these tasks, we leverage existing Korean datasets as well as English datasets. We align the English datasets with Korean using translation and TTS synthesis, filtering out low-quality samples based on CER or human verification to ensure sample quality. In addition, we incorporate listening questions from the korean college scholastic ability test (KCSAT) as authentic long-form Korean speech samples, which are included in the SQA task.

The two proposed speech faithfulness tasks, speech-aware context question answering (SCA-QA) and position-aware question answering (PA-QA), are motivated by our observation that LALMs often fail to fully leverage speech input when generating responses. Specifically, we evaluate speech

faithfulness from two perspectives. First, SCA-QA examines modality faithfulness by examining whether LALMs fully leverage the speech input or rely solely on the text input. By comparing model responses to a text question with and without a counterfactual speech context, we assess how faithfully LALMs utilize the given speech input. To further reflect Korea-specific knowledge, SCA-QA incorporates crawled Korean cultural domain data covering history, sports, and K-pop. Second, PA-QA assesses positional faithfulness with respect to the position of evidence within long-form speech inputs. To construct the dataset, we identify the location of each answer span in the speech, and measure accuracy at each position. This enables a fine-grained analysis of how faithfully LALMs comprehend information across extended speech inputs.

We conduct extensive experiments across both white-box and black-box models, including Qwen3-Omni (Xu et al., 2025), Gemma-3n (Team, 2025a), GPT-audio (Hurst et al., 2024), and Gemini-flash (Comanici et al., 2025), with varying parameter scales, providing a comprehensive analysis of current LALMs on Korean speech understanding.

- We introduce KoALa-Bench, a comprehensive benchmark for evaluating the Korean speech understanding capabilities of LALMs. *To the best of our knowledge, this is the first universal benchmark dedicated to evaluating Korean speech understanding and faithfulness of LALMs.*
- We propose two novel tasks, SCA-QA and PA-QA, to evaluate the speech faithfulness of LALMs in terms of modality and position, examining whether models fully leverage the given speech input.
- To evaluate Korea-specific knowledge, we construct evaluation datasets from KCSAT and crawled cultural domain content covering K-history, K-sports, and K-pop.

2 Related Works

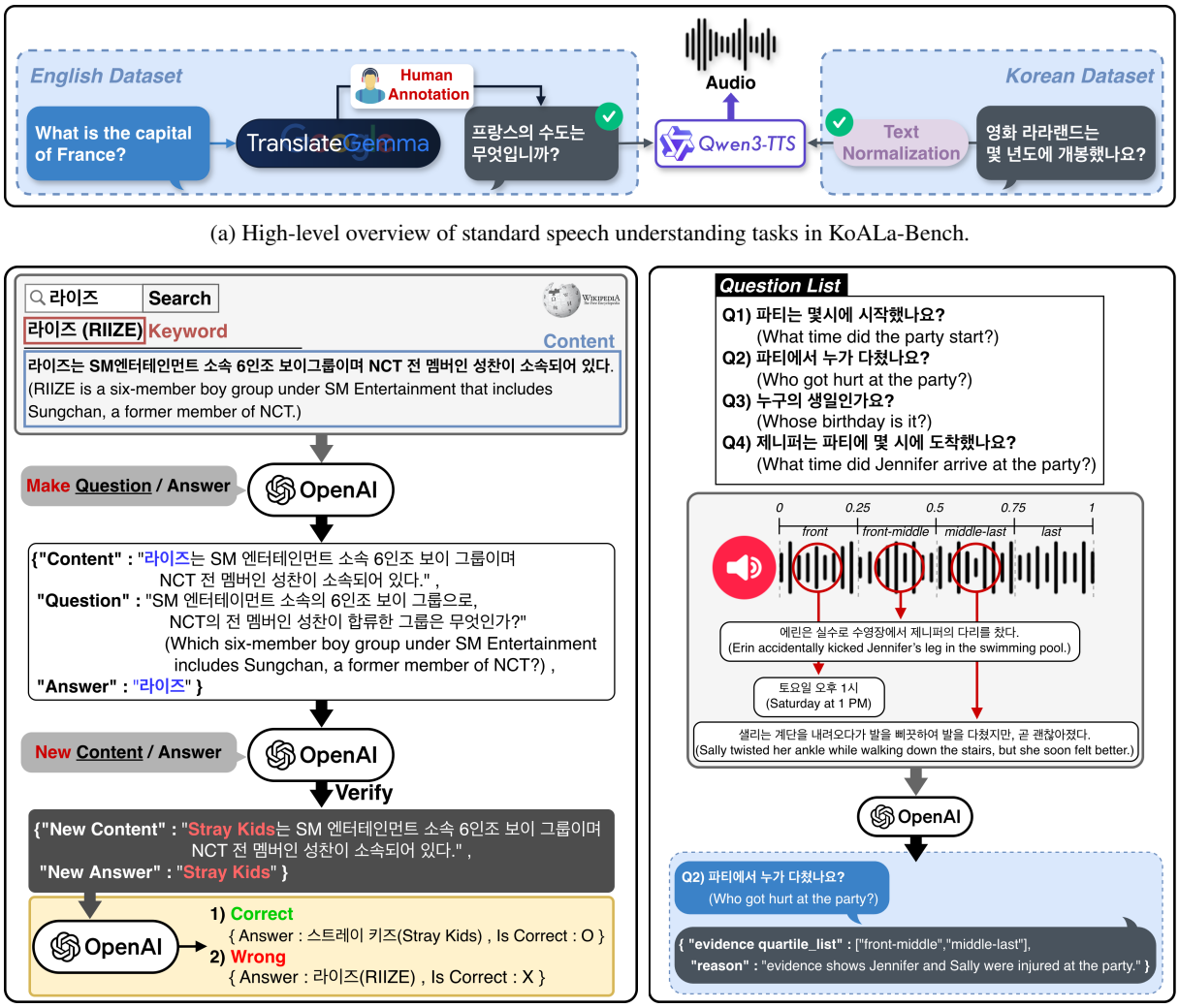
Audio Benchmarks. Recent advances in LALMs have introduced architectures that integrate speech encoders with pretrained LLMs, enabling LLMs to process speech inputs (Team, 2026; Fang et al., 2025; Xu et al., 2025; Chu et al., 2024). As LALMs expand their audio understanding capabilities, various benchmarks have been proposed to evaluate their performance on audio-conditioned tasks. AIR-Bench (Yang et al., 2024) evaluates audio understanding and instruction following across diverse audio inputs, while AudioBench (Wang et al., 2025) provides a comprehensive evaluation suite covering multiple audio tasks. More recent benchmarks explore specialized abilities, including audio dialogue understanding and expert-level reasoning (Gao et al., 2025; Sakshi et al., 2025).

Despite these efforts, existing benchmarks remain largely English-centric, limiting their ability to evaluate the speech understanding capabilities of multilingual LALMs, particularly for languages such as Korean. To address this gap, we propose KoALa-Bench, a unified benchmark designed to evaluate Korean speech understanding and faithfulness of LALMs.

Faithfulness of Multimodal Language Models. Recent multimodal language models (MLLMs) (Team, 2025b; Liu et al., 2025; Tang et al., 2024; Rubenstein et al., 2023) have demonstrated strong capabilities in integrating information from multiple modalities such as text, images, and audio. However, prior studies report that these models often generate responses that are inconsistent with the provided inputs, relying instead on parametric knowledge stored in the language model (Bai et al., 2024b; Liu et al., 2024; Zhai et al., 2023). This phenomenon, commonly referred to as multimodal hallucination, reflects a lack of faithfulness to the given modality. Faithfulness also becomes challenging when models process long speech inputs, as prior studies report that multimodal models often struggle with long-context audio understanding (Ma et al., 2024; Bai et al., 2024a). Despite these observations, the evaluation of faithfulness to speech inputs remains largely underexplored. To facilitate a systematic analysis of this capability, we introduce two complementary evaluation tasks, SCA-QA and PA-QA, designed to assess how faithfully LALMs utilize the given speech input.

3 KoALa-Bench

Overview of KoALa-Bench. We introduce KoALa-Bench, a benchmark designed to assess the speech understanding and speech faithfulness of Korean LALMs. Existing evaluation frameworks primarily focus on generic speech understanding and speech-based QA, with limited consideration



(a) High-level overview of standard speech understanding tasks in KoALa-Bench.

(b) Construction process of the SCA-QA dataset.

(c) The evidence-annotated dataset generation pipeline for PA-QA.

Figure 1: Overview of the KoALa-Bench pipeline. (a) The fundamental pipeline. Synthesizes audio for text-only datasets, translating English text into Korean via Gemma beforehand. (b) The SCA-QA construction process. Generates paired questions (factually aligned vs. intentionally conflicting) to test whether models rely on acoustic context rather than parametric knowledge. (c) The PA-QA dataset generation pipeline. Maps supporting evidence to one of four temporal segments to enable position-aware performance analysis.

of Korean cultural and linguistic characteristics. In addition, prior benchmarks lack evaluation protocols for long duration speech robustness and speech context faithfulness, where speech context faithfulness evaluates whether model predictions are grounded in the acoustic context of the input speech or derived from parametric knowledge. To address these limitations, KoALa-Bench evaluates Korean LALMs across six task dimensions: ASR, ST (English-to-Korean), SQA, SIF, SCA-QA, which measures whether models utilize both speech and text information and PA-QA for assessing long-duration speech robustness. Furthermore, to evaluate model robustness under noisy conditions, we additionally construct noise augmented versions of several datasets by injecting noise into the speech inputs.

3.1 Automatic Speech Recognition (ASR)

Methods. The ASR task evaluates a model's ability to convert spoken utterances into textual transcriptions, serving as a fundamental measure of speech recognition performance. For evaluation, we provide speech inputs to the LALMs and prompt it to generate corresponding Korean transcriptions, which are then compared with ground-truth references.

Datasets. KsponSpeech (Bang et al., 2020) provides spontaneous Korean conversational speech covering diverse everyday topics. Common Voice Korean (Ardila et al., 2020) and Zeroth-Korean[1] contain speech recorded under diverse acoustic conditions as well as relatively clean read speech, enabling evaluation across different speaking styles and recording environments.

Metrics. ASR performance is evaluated using character error rate (CER), which measures the edit distance between predicted transcriptions and ground-truth references at the character level. Before computing CER, we apply a normalization pipeline that removes punctuation and collapses whitespace.

3.2 Speech Translation (ST)

Methods. Korean LALMs should process both English and Korean. Accordingly, KoALa-Bench includes a cross-lingual speech translation task that translates English speech into Korean text. The task evaluates whether the generated output preserves the semantic content of the source speech while producing fluent Korean text. Evaluation is conducted using automatic similarity measures that assess both lexical overlap and semantic consistency between the generated output and the reference translation.

Datasets. We use the ETRI English–Korean Speech Translation Corpus[2], built from TED Talks following the MuST-C[3] data construction methodology. It provides English speech segments paired with Korean translations for evaluation.

Metrics. We evaluate translation quality using BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang* et al., 2020). BLEU and METEOR measure n-gram overlap between the generated translation and the reference, while BERTScore evaluates semantic similarity using contextual embeddings.

3.3 Speech Question Answering (SQA) & Speech Instruction Following (SIF)

Methods. The speech understanding task evaluates the model's ability to answer questions based on acoustic input. In this setting, speech segments are provided as context, which may consist of either natural recordings or speech synthesized using the Qwen3-TTS model (Hu et al., 2026). The model is required to generate responses that correctly follow the textual instruction based on the given speech content. This task evaluates the model's ability to execute text-based instructions conditioned on the provided acoustic context.

3.3.1 SQA

Short-Form Datasets. For the SQA task, we evaluate models on two Korean QA benchmarks: CLIcK and KoBEST-BoolQ. CLIcK is a Korean-native multiple-choice QA benchmark sourced from official exams and textbooks, covering 11 subcategories across language and culture domains. We use the provided evaluation split. KoBEST-BoolQ is the Boolean QA subset of KoBEST, a Korean NLU benchmark built by professional linguists. Each instance consists of a Korean Wikipedia passage and a yes/no question. We use the test subset with binary answers.

Long-Form Datasets. We construct LSQA from the KCSAT listening section collected from EBSi[4] to evaluate long-form speech understanding. LSQA consists of speech passages paired with multiple-choice questions. As the original recordings contain multiple questions in a single long audio file, we segment the audio into question-level samples to construct individual instances.

Metrics. For both benchmarks, we evaluate model performance under two scoring approaches. In the logit-based approach, the log-probability of each candidate token is computed, and the choice with the highest log-probability is selected as the model's prediction. In the generation-based approach, the predicted choice is extracted from the generated response via string matching. Accuracy is used as the final evaluation metric for both approaches.

3.3.2 SIF

Datasets. Instructions are provided as audio and the model generates text responses. To construct Korean instruction-following data, we translate existing English datasets (ALPACA (Taori et al., 2023), Vicuna (Chiang et al., 2023), and OpenHermes[5]) and synthesize them into audio. We additionally synthesize instructions from KUDGE (Son et al., 2024).

[1] https://openslr.org/40/
[2] https://epretx.etri.re.kr
[3] https://ict.fbk.eu/must-c/
[4] https://www.ebsi.co.kr/ebs/pot/poti/main.ebs
[5] https://huggingface.co/datasets/teknium/openhermes

Metrics. Instruction-following performance is evaluated using GPT-4o as a judge model, which assesses whether the generated response correctly follows the instruction.

3.4 Speech Context-Aware Question Answering (SCA-QA)

Methods. Motivated by our observation, as illustrated in Appendix (Table 5), that models often ignore speech context, we introduce SCA-QA. The task constructs paired questions inspired by the methodology of DisentQA (Neeman et al., 2023) to assess whether responses rely on acoustic context or parametric knowledge. To construct the dataset, we crawled documents using keywords related to K-pop, K-history, and K-sports and generated question–answer pairs using the GPT API. Key answer entities were then replaced to create intentionally incorrect answers conflicting with world knowledge, yielding paired questions for the same speech context. All samples were verified by human annotators. SCA-QA consists of paired questions per speech context, one aligned with world knowledge and the other intentionally conflicting, enabling analysis of whether responses rely on acoustic context or parametric knowledge. The full construction pipeline and prompts used at each stage are provided in Appendix A.10.2.

Datasets. SCA-QA is constructed from speech contexts derived from documents crawled using keywords related to K-pop, K-history, and K-sports. Each sample contains paired questions generated via the GPT API and verified by human annotators, where one aligns with common world knowledge and the other intentionally contradicts it through entity replacement.

Metrics. We evaluate model responses using Speech Context Faithfulness (SCF), which measures whether predictions rely on the provided speech context rather than parametric knowledge. For each sample, we first identify cases where the model answers the textual question correctly, and then evaluate the same question with the corresponding speech context. If the model changes its answer according to the speech content when the contextual information conflicts, the response is considered context-faithful. The final score is computed as the proportion of such context-faithful samples. We provide the formal definition in Appendix A.6.

3.5 Postion Aware Question Answering (PA-QA)

Methods. Existing speech benchmarks largely rely on short utterances and lack supporting evidence annotations for each QA pair, making it difficult to analyze which part of the speech context a question depends on. To address this limitation, KoALa-Bench introduces a position-aware evaluation setup by identifying supporting evidence sentences for each question. We construct an evidence annotation pipeline using GPT-4.1-nano to detect supporting sentences and map them to their relative positions within the speech context. Each context is normalized to the range [0,1] and divided into four segments: front (0–0.25), front-middle (0.25–0.5), middle-late (0.5–0.75), and late (0.75–1.0). This dataset enables position-aware analysis of model performance across long speech inputs. The full evidence grounding procedure is described in Appendix A.9.

Datasets. We construct the dataset from MCTest (Richardson et al., 2013), a machine comprehension dataset of short stories and multiple-choice questions. Speech contexts are synthesized from textual passages using Qwen3-TTS (Hu et al., 2026), and supporting evidence sentences are identified with GPT-4.1-nano. After filtering invalid samples, the final dataset contains 327 instances.

Metrics. We follow the same evaluation protocol as in SQA and report accuracy. Each question is associated with a normalized position in the speech context, enabling position-aware analysis.

4 Experiments

4.1 Evaluation setup

We evaluated five models, including Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025), Google-gemma-3n-E4B-it (Team, 2025a), Voxtral-Mini-3B-2507 (Liu et al., 2025), GPT-audio-mini[6], and Gemini-flash-lite (Comanici et al., 2025). For open-weight models, inference was performed using vLLM (Kwon et al., 2023) on two NVIDIA A6000 Blackwell GPUs with 96 GiB memory each. For proprietary models, we used the official API. The datasets used for evaluation are summarized in Table 1.

To assess robustness under acoustic degradation, we additionally evaluated models on noise-

[6]https://developers.openai.com/api/docs/models/gpt-audio-mini

Dataset		Metrics	Noise	# Samples	Hours(h)	Length(s)		
						Avg	Min	Max
ASR								
Zeroth		CER ↓	✓	457	1.191	9.381	4.558	20.453
Common Voice			✓	523	0.726	4.998	1.728	12.564
KsponSpeech-Clean			✗	3000	2.644	3.173	0.286	19.357
KsponSpeech-Other			✗	3000	3.803	4.563	0.399	20.093
ST								
ETRI-TST-Common		BLEU / METEOR / BERTScore ↑	✗	2532	4.877	650.251	210.233	1417.741
ETRI-TST-HE			✗	544	2.282	746.788	551.358	1117.391
SQA								
Short-Form	CLIcK	Accuracy ↑	✓	1315	4.180	11.444	1.600	275.402
	KoBest BoolQ		✓	1404	11.950	30.642	4.640	88.308
Long-Form	KCSAT		✓	82	2.919	128.182	62.976	208.020
SIF								
KUDGE		Score (GPT as Judge) ↑	✓	557	1.863	10.365	3.440	23.497
Vicuna			✓	70	0.192	8.623	2.640	16.297
OpenHermes			✓	78	0.174	6.450	1.680	18.617
Alpaca			✓	69	0.111	4.375	1.760	9.017
SCA-QA								
K-history (Before Chosun)		SCF Score ↑	✓	101	0.815	29.053	11.280	61.280
K-history (After Chosun)			✓	82	0.997	43.780	3.520	100.000
K-sports			✓	88	1.698	69.488	20.400	92.640
K-pop			✓	103	1.457	50.926	8.000	82.400
PA-QA								
MCTest		Accuracy ↑	✓	325	8.732	96.142	38.400	205.120

Table 1: Datasets used in the proposed KoALa-Bench.

augmented samples constructed by injecting background noise from the SELECTSTAR noise dataset[7], which contains diverse environmental sounds such as traffic and ambient location noise. Further details on the noise sources and the augmentation procedure are provided in the Appendix A.2.

In terms of the evaluation protocol, since the models under evaluation incorporate large language models as their backbone, they can be highly sensitive to prompt phrasing (Maia Polo et al., 2024; Sclar et al., 2024; Zhuo et al., 2024). To account for this, we evaluated each task using four Korean prompts expressing the same instruction with slight lexical and structural variations, and reported the best result as the final score. The full prompt list is provided in Appendix A.10.1.

4.2 Main Results

4.2.1 Standard Speech Understanding

Table 2 presents the overall performance of various LALMs across ASR, ST, SQA, and SIF tasks. Overall, Qwen3 shows the strongest performance on speech understanding tasks, achieving the lowest CER across all ASR benchmarks and the highest BERTScore on ST datasets. It also performs competitively on spoken QA, particularly on more challenging benchmarks such as CLIcK and KCSAT. In contrast, GPT achieves the highest GPTScore on SIF tasks, indicating stronger instruction-following capabilities. Across most tasks, Qwen3 and GPT show relatively small performance drops under noisy conditions, suggesting stronger robustness to noisy speech inputs compared to other models.

4.2.2 Speech Faithfulness

SCA-QA. As shown in Table 3, GPT-audio-mini achieved the highest Text-only accuracy across domains, indicating strong Korea-specific parametric knowledge. However, under counterfactual speech context, it tended to follow its prior knowledge rather than the speech input, suggesting limited use of the speech modality. Qwen3 maintained high Text-only accuracy while exhibiting the lowest prediction preservation, indicating the most effec-

[7] https://open.selectstar.ai/ko/cochl

Dataset		Metric	Qwen3-omni	Gemma-3n	GPT-audio	Voxtral	Gemini-flash
ASR							
Zeroth		CER ↓	**3.33** / **3.91**	100↑ / 100↑	6.87 / 9.00	40.92 / 39.07	13.60 / 14.56
Common Voice			**4.96** / **6.78**	100↑ / 100↑	33.05 / 36.21	60.10 / 58.98	13.74 / 26.74
KsponSpeech-Clean			**8.46**	100↑	100↑	62.62	83.19
KsponSpeech-Other			**7.91**	100↑	63.64	56.04	45.14
ST							
ETRI-TST-COMMON		BERTScore ↑	**93.40**	87.39	93.10	92.73	91.60
ETRI-TST-HE			**93.96**	87.79	93.69	93.09	92.17
SQA							
Short-Form	CLIcK	Accuracy ↑	64.04 / 62.30	35.79 / 35.71	61.64 / 60.07	42.58 / 42.92	**67.27** / **66.69**
	KoBEST BoolQ		51.34 / 51.16	50.89 / 50.54	51.88 / 50.45	50.54 / 50.54	**52.86** / **54.92**
Long-Form	KCSAT		**83.53** / **84.71**	34.12 / 40.00	52.90 / 47.10	69.41 / 72.94	81.18 / 78.82
SIF							
KUDGE		Score (GPT as Judge) ↑	71.87 / 71.82	71.38 / 70.69	**74.07** / **73.99**	61.97 / 61.70	70.29 / 70.39
Vicuna			79.64 / 78.43	80.21 / 80.00	**82.14** / **81.79**	67.79 / 69.50	76.43 / 73.29
OpenHermes			86.54 / 85.19	84.62 / 85.96	**89.42** / **89.62**	69.10 / 69.62	82.69 / 80.71
Alpaca			84.06 / 83.04	82.97 / 83.36	**90.58** / **90.58**	72.90 / 72.46	85.94 / 86.88

Table 2: Comparison across ASR, ST, SQA, and SIF. ASR, SQA, and SIF are evaluated on clean and noisy audio. KsponSpeech-Clean/Other denote subsets. 100 ↑ indicates CER > 100%. Blue indicates noisy results.

Dataset		Metric	Qwen3-omni	Gemma-3n	GPT-audio	Voxtral	Gemini-flash
SCA-QA							
K-history (Before)		Text-only Accuracy ↑	55.45	48.51	69.30	41.58	**79.21**
K-history (After Chosun)			79.27	71.95	79.30	51.22	**87.80**
K-sports			55.68	53.41	69.30	27.27	**78.41**
K-pop			67.96	68.93	69.90	42.72	**89.32**
K-history (Before)		SCF Score ↑	**94.64**	67.35	61.40	85.71	66.25
K-history (After Chosun)			**92.31**	45.76	32.30	85.71	59.72
K-sports			93.88	76.60	39.30	**95.83**	86.96
K-pop			**95.71**	73.24	37.50	88.64	78.26
PA-QA							
MC-Test Total		Accuracy ↑	**93.54** / **93.85**	48.92 / 48.92	77.23 / 79.69	84.92 / 86.15	92.31 / 92.00
Position - Aware	Front		91.93 / 91.93	45.34 / 44.72	72.05 / 75.78	86.34 / 85.09	**95.03** / **93.17**
	Front-Middle		**96.70** / **94.51**	51.65 / 53.85	83.52 / 83.52	82.42 / 85.71	89.01 / 87.91
	Middle-Last		**87.80** / **87.80**	47.56 / 47.56	74.39 / 70.73	76.83 / 81.71	82.93 / 85.37
	Last		**94.12** / **98.04**	45.10 / 45.10	80.39 / 86.27	82.35 / 82.35	92.16 / 98.04

Table 3: Performance comparison on speech faithfulness tasks. SCA-QA reports text-only accuracy and SCF score. PA-QA reports accuracy by evidence position. Blue indicates results on the dataset with additional background noise.

tive use of speech input. Voxtral showed relatively stronger speech faithfulness despite lower Text-only accuracy. We further evaluated models using correct speech inputs to verify modality alignment. Additional analysis is provided in the Analysis.

PA-QA. On the MC-Test benchmark (Table 3), Qwen3-Omni achieved the highest accuracy under both clean and noisy conditions, with particularly stronger performance under noise. Gemma-3n exhibited a monotonic decline in accuracy across the positional analysis, indicating a stronger reliance on early speech information. GPT-audio-mini performed particularly poorly in the Front-Middle region, whereas Voxtral showed weaknesses in the Front-Middle and Late segments.

5 Analysis

5.1 Speech Faithfulness

SCA-QA. To further assess speech-modality alignment, we compared model predictions under wrong and correct speech contexts. Table 4 summarizes the possible prediction patterns and their interpretations. Case 1 indicated that the model followed the speech input, whereas Case 3 suggested that the speech context was ignored. In contrast, Cases 2 and 4 corresponded to misalignment, where the model failed to produce the correct answer even under the correct speech context. Figure 2 shows that

	Case 1	Case 2	Case 3	Case 4
Speech Context				
Wrong	O	O	X	X
Correct	O	X	O	X
Outcome				
Aligned	✓	✗	✓	✗
Speech Utilization	✓	–	✓	–

Table 4: Case analysis for speech-modality alignment. Case categorization is performed only for instances where the text-context condition is correctly answered.

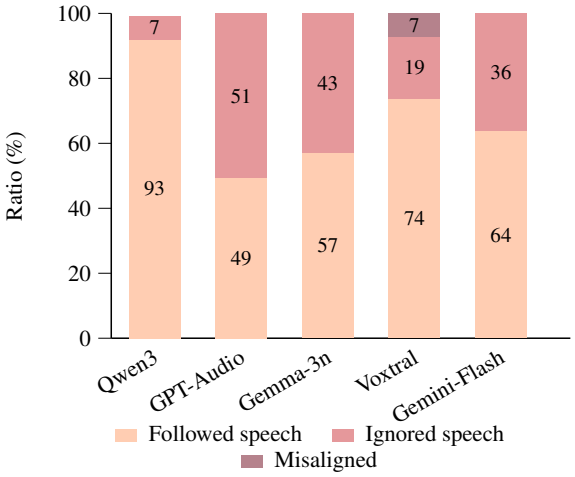
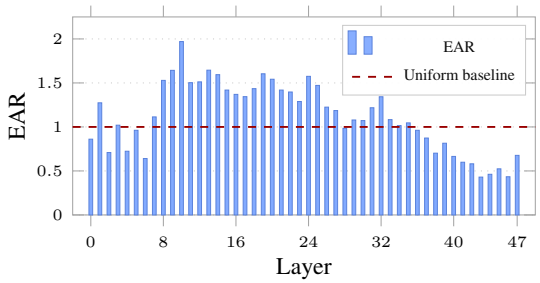


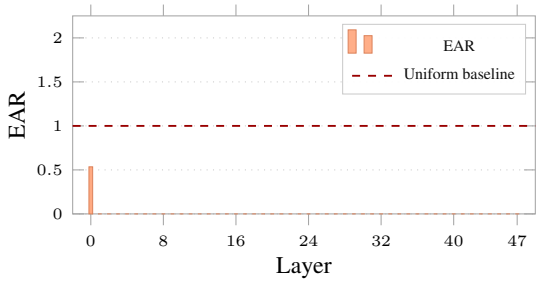
Figure 2: Model behavior under correct speech contexts on the SCA-QA benchmark. The bar chart reports results for the K-history (before Chosun) subset.

Voxtral exhibited speech–text misalignment. This result indicates that KoALa-Bench enables analysis of speech–text alignment failures in LALMs.

PA-QA. To analyze whether models attend to relevant speech evidence, we computed the Evidence Attention Ratio (EAR) for evidence and non-evidence sentences. The formal definition of EAR is provided in Appendix A.7. As shown in Figure 3 (a), EAR for evidence sentences was mostly above 1 across layers, indicating higher attention to ground-truth evidence spans. In contrast, Figure 3 (b) showed that EAR for irrelevant sentences was consistently well below 1 across all layers, indicating suppressed attention to non-evidence regions. These results show that models can distinguish and focus on semantically relevant evidence in speech. This analysis was enabled by explicit evidence annotations in PA-QA, allowing fine-grained examination of attention behavior. A qualitative example is provided in Appendix A.11. Importantly, these findings suggest that models attend to the intended evidence positions, indicating that PA-QA's position-based evaluation can meaningfully reflect model behavior for assessing speech faithfulness.



(a) Evidence sentence



(b) Irrelevant sentence

Figure 3: Per-layer EAR for (a) the evidence sentence and (b) an irrelevant sentence.

6 Conclusion

In this paper, we introduce KoALa-Bench, a benchmark for evaluating Korean speech understanding in audio language models. KoALa-Bench includes ASR, ST, SQA, SIF, and two speech faithfulness tasks, SCA-QA and PA-QA, which evaluate modality and positional faithfulness to speech input with dedicated metrics. To assess Korea-specific knowledge, we incorporate KCSAT and crawled Korean-culture datasets. We evaluate both white-box and black-box models under clean and noisy conditions.

7 Limitations

While KoALa-Bench provides a comprehensive evaluation framework for Korean speech understanding, it has two limitations. First, the benchmark is currently limited to Korean, leaving other non-English languages unexplored. Nevertheless, we believe our benchmark serves as an important step toward multilingual speech evaluation beyond English. Second, while the ASR task data and KCSAT listening questions consist of real human speech, the remaining datasets rely on TTS-synthesized speech, which may not fully capture

the diversity of natural spontaneous speech. We leave the extension to real-world speech data and other languages as future work.

8 Ethical Considerations

We construct KoALa-Bench from publicly available data while complying with applicable copyright and licensing requirements. All data are used for non-commercial academic research, and we do not redistribute raw copyrighted content, providing only processed annotations and metadata when permitted. The dataset may inherit cultural and domain-specific biases from the source data, and evaluated models may exhibit hallucinations or unintended behaviors. This benchmark is intended for analyzing model behavior, particularly reliance on speech context, and is not designed for safety-critical use. We encourage responsible use and transparent reporting of model limitations.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2024a. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 3119–3137.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024b. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Jeong-Uk Bang, Seung Yun, Seung-Hi Kim, Mu-Yeol Choi, Min-Kyu Lee, Yeo-Jeong Kim, Dong-Hyun Kim, Jun Park, Young-Jik Lee, and Sang-Hun Kim. 2020. Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition. *Applied Sciences*, 10(19):6936.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. 2025. LLaMA-omni 2: LLM-based real-time spoken chatbot with autoregressive streaming speech synthesis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18617–18629, Vienna, Austria. Association for Computational Linguistics.

Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. 2025. Benchmarking open-ended audio dialogue understanding for large audio-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4763–4784.

Jung-Woo Ha, Kihyun Nam, Jingu Kang, Sang-Woo Lee, Sohee Yang, Hyunhoon Jung, Eunmi Kim, Hyeji Kim, Soojin Kim, Hyun Ah Kim, and 1 others. 2020. Clovacall: Korean goal-oriented dialog speech corpus for automatic speech recognition of contact centers. *arXiv preprint arXiv:2004.09367*.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, and 1 others. 2026. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Alexander H Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy,

and 1 others. 2025. Voxtral. *arXiv preprint arXiv:2507.13264*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations*, volume 2024, pages 57689–57733.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, and 1 others. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson de Oliveira, Yuekai Sun, and Mikhail Yurochkin. 2024. Efficient multi-prompt evaluation of llms. *Advances in Neural Information Processing Systems*, 37:22483–22512.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, and 1 others. 2023. Audiopalm: A large language model that can speak and listen.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Guijin Son, Hyunwoo Ko, Hoyoung Lee, Yewon Kim, and Seunghyeok Hong. 2024. Llm-as-a-judge & reward model: What they can and cannot do. *arXiv preprint arXiv:2409.11239*.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team. 2025a. Gemma 3n.

Meituan LongCat Team. 2025b. Longcat-flash-omni technical report. *Preprint*, arXiv:2511.00279.

NAVER Cloud HyperCLOVA X Team. 2026. Hyperclova x 8b omni. *arXiv preprint arXiv:2601.01792*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy Chen. 2025. 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)*, pages 4297–4316.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and 1 others. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998.

Zeroth Project. 2018. Zeroth-korean: Open korean speech corpus. https://www.openslr.org/40/. OpenSLR SLR40.

Bohan Zhai, Shijia Yang, Chenfeng Xu, Sheng Shen, Kurt Keutzer, Chunyuan Li, and Manling Li.

2023. Halle-control: controlling object hallucination in large multimodal models. *arXiv preprint arXiv:2310.01779*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. Prosa: Assessing and understanding the prompt sensitivity of llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976.

A Appendix

A.1 Failure to Leverage Speech Context

Field	Content
Speech Context	K/DA는 비디오 게임이던 "리그 오브 레전드"의 챔피언 아리, 이블린, 카이사, 아칼리으로 구성된 가상의 여성 걸 그룹이다. K/DA라는 이름은 Kill/Death/Assist에서 영어의 한자 씩의 이름으로 왔으며, 팬덤 이름은 블레이즈이다. *(K/DA is a virtual female group of LoL champions Ahri, Evelynn, Kai'Sa, and Akali. The name derives from Kill/Death/Assist, and the fandom name is Blaze.)*
Question	팬덤 이름이 블레이즈인 그룹은? *(Which group's fandom name is Blaze?)*
Answer	(D) K/DA
Model Response *w/ question-only*	(A) ITZY
Model Response *w/ Speech+question*	(A) ITZY

Table 5: A case from Qwen3-Omni-30B-A3B-Instruct where the model answers incorrectly given the question only. Even when the speech context explicitly contains the correct answer, the model still produces the same incorrect response, indicating a failure to leverage the provided speech context.

A.2 Details of Noise Injection for Robustness Experiments

Dataset	# Samples
Traffic	6,292
Environmental	13,300
Other	8,000

Table 6: Types of Environmental Noise Used for Noise Augmentation.

To evaluate robustness under acoustic degradation, we augmented the speech samples using both real environmental noise from the SELECTSTAR noise dataset and additional synthetic noise. The SELECTSTAR dataset contains diverse environmental sounds such as traffic noise and ambient location noise and the detailed noise information is provided in Table 6. In addition to real-world noise, we also introduced synthetic noise types, including white noise, pink noise, brown noise, electrical hum, and impulse noise, to simulate a wider range of acoustic disturbances. All noise signals were mixed with speech at randomly sampled signal-to-noise ratio (SNR) levels to generate the noise-augmented evaluation samples.

A.3 Additional Speech Translation Metrics

Dataset	ETRI-TST-COMMON		ETRI-TST-HE	
	BLEU	METEOR	BLEU	METEOR
Qwen3-omni	**28.53**	**52.02**	**31.71**	**55.79**
Gemma-3n	5.17	16.97	5.24	17.50
GPT-audio	26.06	49.37	28.98	52.96
Voxtral	23.07	45.95	26.08	50.16
Gemini-flash	22.88	47.15	26.95	52.59

Table 7: Translation performance on ETRI evaluated using BLEU and METEOR scores. Bold indicates the best result.

We additionally provide translation evaluation results using BLEU and METEOR. As shown in Table 7, Qwen3-Omni achieved the best performance on both ETRI-COMMON and ETRI-HE, obtaining the highest BLEU and METEOR across the compared models.

A.4 CER Normalization Rules

We evaluate Korean ASR using character error rate (CER) with a normalization pipeline tailored to the linguistic properties of Korean numerals.

Motivation. In Korean ASR evaluation, surface-level discrepancies between the ground-truth transcription and the model hypothesis frequently arise due to punctuation and spacing differences. ASR models often omit or insert punctuation that is present (or absent) in the ground truth, and Korean word spacing conventions are inherently ambiguous—even native speakers segment words inconsistently. Naïvely computing CER on such pairs would incur false penalties despite semantically identical transcriptions. To address this, we apply a normalization pipeline that removes punctuation and whitespace from both the reference and the hypothesis before CER computation.

Pipeline. Before computing CER, we apply a normalization step to both the reference and the hypothesis, where all punctuation, special characters, and whitespace are removed, retaining only Hangul characters and Arabic digits. This prevents surface-level discrepancies, such as punctuation insertion or omission and spacing inconsistencies, from inflating CER, ensuring that semantically identical transcriptions are not unfairly penalized. Table 8 shows concrete examples covering the normalization rules.

Ground Truth (GT)	Prediction	CER		Explanation
Punctuation & space normalization — Symbols removed, spaces collapsed				
안녕하세요!	안녕하세요	0%	3	Punctuation "!" removed from GT
정말요?	정말요	0%	3	Question mark "?" removed from GT
네, 알겠습니다.	네알겠습니다	0%	3	Comma and period removed
아... 그래요	아그래요	0%	3	Ellipsis and spaces removed
감사 합니다	감사합니다	0%	3	Spaces removed from both sides
Hangul ground truth — No number conversion applied				
오백 원	오백원	0%	3	Spaces removed; Hangul preserved as-is
오백 원	오천원	> 0%	×	No variant generated; "오백" ≠ "오천"
세 잔	세잔	0%	3	Spaces removed; Hangul preserved as-is
세 잔	삼잔	> 0%	×	No variant generated for Hangul GT

Table 8: Examples of CER evaluation under the normalization pipeline. Group 1 shows punctuation and space handling; group 2 shows that Hangul ground truth is preserved as-is. 3 = accepted (CER = 0%), × = rejected (CER > 0%).

A.5 CER Computation

CER is computed at the corpus level via micro-averaging:

$$\text{CER} = \frac{\sum_{i=1}^{N} d(r_i,\, h_i)}{\sum_{i=1}^{N} |r_i|} \tag{1}$$

where $d(r_i, h_i)$ is the Levenshtein edit distance between the normalized reference r_i and hypothesis h_i, and $|r_i|$ is the character length of the normalized reference.

A.6 Evaluation Metric of SCA-QA

In this section, we formally define SCF-score, which measures whether model predictions rely on the provided speech context rather than parametric knowledge.

Let $\mathcal{D} = \{(q_i, a_i^{\text{text}}, a_i^{\text{speech}}, s_i)\}_{i=1}^{N}$ denote the evaluation set, where q_i is a textual question, a_i^{text} is the ground-truth answer to q_i, a_i^{speech} is the answer derived from the speech context s_i that conflicts with a_i^{text}, and f denotes the model under evaluation. We define

- $\hat{a}_i^{\text{text}} = f(q_i)$: the model's prediction given only the textual question,
- $\hat{a}_i^{\text{speech}} = f(q_i, s_i)$: the model's prediction given both the question and the speech context.

We first identify the subset of samples where the model answers the textual question correctly as

$$\mathcal{D}_{\text{correct}} = \left\{ i \mid \hat{a}_i^{\text{text}} = a_i^{\text{text}} \right\}, \tag{2}$$

and then compute SCF as the proportion of these samples for which the model faithfully follows the speech context:

$$\text{SCF} = \frac{\left| \left\{ i \in \mathcal{D}_{\text{correct}} \mid \hat{a}_i^{\text{speech}} = a_i^{\text{speech}} \right\} \right|}{|\mathcal{D}_{\text{correct}}|}. \tag{3}$$

A higher SCF indicates that the model is more faithful to the provided speech context rather than relying on its parametric knowledge.

A.7 Evidence Attention Ratio (EAR)

To analyze whether models attend to the relevant speech evidence, we measure the Evidence Attention Ratio (EAR), which quantifies how strongly the model focuses on the ground-truth evidence span relative to what would be expected under a uniform attention distribution.

Let E denote the set of tokens belonging to the ground-truth evidence span, N denote the set of non-evidence tokens within the speech context (i.e., $N = S - E$, where S is the full set of speech tokens), and a_i denote the attention weight assigned to token i from a given query token. We first compute the raw attention ratio as

$$\text{EAR}_{\text{raw}} = \frac{\sum_{i \in E} a_i}{\sum_{i \in E} a_i + \sum_{j \in N} a_j}. \tag{4}$$

Under a uniform attention distribution, the expected value of EAR_{raw} equals the proportion of evidence tokens:

$$\text{EAR}_{\text{uniform}} = \frac{|E|}{|E| + |N|}. \tag{5}$$

We define EAR as the ratio of the raw attention ratio to this uniform baseline:

$$\mathrm{EAR} = \frac{\mathrm{EAR}_{\mathrm{raw}}}{\mathrm{EAR}_{\mathrm{uniform}}}. \tag{6}$$

An EAR value of 1.0 indicates that the model distributes attention uniformly across speech tokens, while values greater than 1.0 indicate above-uniform attention to the evidence span. This normalization ensures that EAR is comparable across samples with different evidence lengths.

For each example, we compute EAR from the attention weights of the last generated token to all speech tokens, averaged across attention heads. We then report the layer-wise EAR to examine how evidence attention evolves across model depth.

A.8 SCA-QA Pipeline

Algorithm 1: SCA-QA Dataset Construction Pipeline

Input: Wikipedia passages $\mathcal{D}$, LLM client $\mathcal{G}$, TTS engine $\mathcal{T}$
Output: Multiple-choice QA samples with counterfactual speech contexts
Stage 1: Question and Answer Generation;
for *passage* $x \in \mathcal{D}$ **do**
  Generate question q and answer a from x using $\mathcal{G}$ with 3-shot prompting;
Stage 2: Entity Replacement;
for *each* (x, q, a) **do**
  Generate a new entity a' of the same type as a using $\mathcal{G}$, ensuring $a' \notin x$;
Stage 3: Sentence Transformation;
for *each* (x, a, a') **do**
  Replace all occurrences of a with a' in x to produce counterfactual context x' using $\mathcal{G}$;
Stage 4: TTS Text Normalization;
for *each* x' **do**
  Normalize numerals and non-Korean characters into spoken Korean forms using $\mathcal{G}$;
  Synthesize speech from the normalized text using $\mathcal{T}$;
Stage 5: Distractor Generation;
for *each* (a, a') **do**
  Generate two distractors $\{d_1, d_2\}$ of the same entity type using $\mathcal{G}$;
  Construct choices $\mathcal{C} = \{a', a, d_1, d_2\}$ and shuffle; record correct index;
Stage 6: Answer Verification;
for *each* $(x', q, \mathcal{C}, a')$ **do**
  Query $\mathcal{G}$ with x', q, and $\mathcal{C}$;
  Discard sample if $\mathcal{G}$'s response $\neq a'$;

A.9 Two-Stage Evidence Grounding

Algorithm 2: Two-Stage Evidence Grounding: Text $\rightarrow$ Audio

Input: Samples $\mathcal{D}$, GPT client $\mathcal{G}$, ASR model $\mathcal{W}$
Output: Samples with evidence spans, validity, and audio time spans
Stage 1: Text evidence grounding;
for $r \in \mathcal{D}$ **do**
    Extract context x, question q, and gold answer $a^\star$;
    if $x = \emptyset$ **then**
        Set `is_valid` $\leftarrow$ `False`; continue;
    Split x into sentences $\{s_i\}$ with text spans;
    Select candidate evidence sentences $\mathcal{C}$ using rule-based matching and GPT fallback;
    Verify $\mathcal{C}$ with GPT (mandatory for non-empty x), and re-verify final $\mathcal{C}$ once;
    Set `is_valid` $\leftarrow$ (`supported` $\land \neg$`need_more`);
    Save evidence sentences and their text spans (and text-based positions);
Stage 2: Audio evidence grounding;
$\mathcal{D}_v \leftarrow \{r \in \mathcal{D} \mid r.$`is_valid` $=$ `True`$\}$;
for $r \in \mathcal{D}_v$ **do**
    Transcribe the audio with word-level timestamps using $\mathcal{W}$;
    Build a text-to-time mapping from the transcript;
    for *each evidence sentence* s **do**
        Align s to the transcript (exact, then fuzzy fallback);
        Convert the aligned text span into an audio time span $[t_s, t_e]$; if unmatched, set `None`;
    Save the audio time spans for the evidence sentences;

A.10 Prompts Lists

A.10.1 Evaluation Prompts

Prompts for KoALa-Bench Evaluation

ASR (Automatic Speech Recognition)
Objective: Transcribe the given speech into Korean text.

Prompts:
1) 주어진 음성을 한국어로 전사하고 전사된 한국어 텍스트만 출력해 주세요.
2) 다음 음성을 한국어로 전사해 주시고 설명이나 추가 문장 없이 전사 결과만 제공해 주세요.
3) 제공된 음성의 내용을 한국어로 정확하게 전사해 주시고 전사된 문장만 출력해 주세요.
4) 맥락으로 주어진 음성을 한국어로 전사해 주시되 다른 설명 없이 전사된 한국어 텍스트만 응답해 주세요.

ST (Speech Translation)
Objective: Translate the given speech into Korean.

Prompts:
1) 이 오디오를 한국어로 번역해 주세요. 번역 결과만 출력하고 다른 텍스트는 출력하지 마세요.
2) 다음 문장을 의미를 변경하지 말고 자연스러운 한국어로 번역하시오.
3) 다음 영어 문장을 정보의 추가나 삭제 없이 정확하게 한국어로 번역하시오.
4) 다음 문장의 의미를 그대로 유지한 채 한국어로 번역하고 번역 결과만 출력하시오.

SQA / PA-QA
Objective: Answer questions based on the given speech.

Prompt Prefix:
1) <One-shot prompt> 다음 음성을 듣고 질문에 맞는 답을 고르세요.
2) <One-shot prompt> 주어진 음성을 듣고 가장 적절한 답을 선택하세요.
3) <One-shot prompt> 음성을 잘 듣고 알맞은 답을 골라 주세요.
4) <One-shot prompt> 아래 음성의 내용을 참고하여 올바른 답을 선택하세요.

Answer Suffix: \n답:

One-shot prompt:
예시:\n질문: 철수는 어디에서 친구를 만났나요?\n(A) 학교 (B) 공원 (C) 도서관 (D) 카페\n답: (B)\n\n

SIF (Speech Instruction Following)
Objective: Assess the model's ability to follow spoken instructions.

Prompts:
1) 음성에서 제시된 질문을 참고해 주시기 바랍니다.
2) 음성의 정보를 활용하여 다음 질문에 답해 주세요.
3) 음성을 듣고 다음 질문에 답해 주세요.
4) 음성 내용을 바탕으로 다음 질문에 답해 주세요.

SCA-QA (Speech Context Aware Question Answering)
Objective: Answer questions based solely on the text question, comparing speech-grounded responses.
A one-shot example is used to elicit answers in a consistent, parseable format.

Prompt (Text-only):
1) <One-shot example> 다음 질문에 답변하세요.
2) <One-shot example> 질문에 맞는 답을 고르세요.
3) <One-shot example> 아래 질문에 가장 적절한 답을 선택하세요.
4) <One-shot example> 주어진 선택지 중 올바른 답을 고르세요.

Prompt (with speech context):
1) <One-shot prompt> 음성에서 들려준 내용만을 근거로 다음 질문에 답변하세요.
2) <One-shot prompt> 음성에서 제공된 정보에만 기반하여 질문에 맞는 답을 고르세요.
3) <One-shot prompt> 반드시 음성에서 들은 내용을 바탕으로 아래 질문에 가장 적절한 답을 선택하세요.
4) <One-shot prompt> 음성의 내용만을 참고하여 주어진 선택지 중 올바른 답을 고르세요.

One-shot prompt:
예시:
질문: 철수는 어디에서 친구를 만났나요?
(A) 학교 (B) 공원 (C) 도서관 (D) 카페
답: (B)

A.10.2 SCA-QA Dataset Construction Prompts

Stage 1: Question and Answer Generation

Objective: Generate a factual question and a short answer from the given context.

문단(context)을 읽고, 그 내용으로 답할 수 있는 자세한 질문을 만든 뒤, 정답을 한 단어(또는 짧은 구)로 써 주세요.
- 질문: 문단에 나온 사실을 묻는 구체적인 질문. 질문 문장 안에 정답 단어를 넣지 마세요.
- 답변: 문단에서 묻는 대상(인물, 사건, 장소, 작품명 등)을 짧게 한 단어로.

[예시 1]
context: 살수 대첩은 제2차 고수 전쟁을 고구려의 승리로 장식한 전투이다. 평양을 직공했던 수나라의 30만 군대가 살수를 건너 회군하던 중 을지문덕 군의 공격을 받아 궤멸적인 피해를 입고 패주했다. 귀주 대첩, 한산도 대첩과 함께 한국사 3대 대첩 중 하나로 불린다.
질문: 제2차 고수 전쟁을 고구려의 승리로 장식한 전투이며, 수나라 30만 군대가 을지문덕의 공격을 받아 궤멸적 피해를 입고 패주한, 한국사 3대 대첩 중 하나로 불리는 전투는 무엇인가?
답변 : 살수 대첩

[예시 2]
context: 안시성 전투는 제1차 고당 전쟁 중 당 태종, 이세적이 지휘하는 군대가 대규모로 고구려를 침공하여 2개월 가량 안시성을 포위하고 공격했으나, 끈질긴 항전으로 이를 물리치고 고구려가 승리한 싸움이다.
질문: 제1차 고당 전쟁 중 당 태종과 이세적이 이끄는 군대가 2개월 동안 포위·공격했으나 고구려가 끈질긴 항전 끝에 승리한 전투는 무엇인가?
답변 : 안시성 전투

[예시 3]
context: 황산벌 전투는 660년 8월 20일에서 8월 21일 황산벌에서 신라 군과 백제 군 사이에 일어났던 전투이다. 삼국유사 태종무열왕조, 삼국사기 계백 열전 등에 전투 내용이 나온다.
질문: 660년 8월 20일과 21일 사이에 황산벌에서 신라 군과 백제 군 간에 벌어진 전투로, 삼국유사와 삼국사기 계백 열전에 기록된 전투의 이름은 무엇인가?
답변 : 황산벌 전투

–
아래 문단을 보고 같은 형식으로 질문과 답변을 작성하세요. (질문: ... / 답변 : ...)
context:
{context}

질문:

Stage 2: Entity Replacement

Objective: Find a different entity of the same type to replace the original answer.

컨텍스트와 현재 엔티티가 주어집니다. 이 엔티티와 **같은 유형**의 **다른** 엔티티 하나만 골라서 알려주세요. (예: 역사 전투 → 다른 전투명, 걸그룹 → 다른 걸그룹명) 문장을 바꾸지 말고, 새 엔티티 이름만 출력하세요.
IMPORTANT!!! 컨텍스트 안에 있는 단어를 새로운 엔티티로 작성하지 마세요. 새로운 엔티티는 문장 안에 없어야 합니다.

[예시 1]
문장: 살수 대첩은 제2차 고수 전쟁 을 고구려 의 승리로 장식한 전투이다. 평양을 직공했던 수나라의 30만 군대가 살수를 건너 회군하던 중 을지문덕 군의 공격을 받아 궤멸적인 피해를 입고 패주했다. 귀주 대첩, 한산도 대첩과 함께 한국사 3대 대첩 중 하나로 불린다.
현재 엔티티: 살수 대첩

→ 새로운 엔티티: 안시성 전투
불가능한 엔티티 : 한산도 대첩(컨텍스트 안에 존재함)

[예시 2]
문장: 크레용팝은 대한민국의 오인조 걸 그룹이다...
현재 엔티티: 크레용팝
→ 새로운 엔티티: 걸스데이

–
현재 엔티티: {answer}

문장:
{text_to_tts}

아래 형식으로만 출력하세요.
새로운 엔티티:

Stage 3: Sentence Transformation

Objective: Replace the original entity with the new entity in the context.

문장에서 '현재 엔티티'를 '새 엔티티'로 **바꾼 문장**을 만들어 주세요.

[예시 1]
문장: 살수 대첩은 제2차 고수 전쟁을 고구려의 승리로 장식한 전투이다. 평양을 직공했던 수나라의 30만 군대가 살수를 건너 회군하던 중 을지문덕 군의 공격을 받아 궤멸적인 피해를 입고 패주했다. 귀주 대첩, 한산도 대첩과 함께 한국사 3대 대첩 중 하나로 불린다.
현재 엔티티: 살수 대첩
새 엔티티: 안시성 전투
→ 수정된 문장: 안시성 전투는 제2차 고수 전쟁을 고구려의 승리로 장식한 전투이다. 평양을 직공했던 수나라의 30만 군대가 살수를 건너 회군하던 중 을지문덕 군의 공격을 받아 궤멸적인 피해를 입고 패주했다. 귀주 대첩, 한산도 대첩과 함께 한국사 3대 대첩 중 하나로 불린다.

[예시 2]
문장: 크레용팝은 대한민국의 오인조 걸 그룹이다. '크레용팝'이라는 이름은 음악이라는 도화지에 멤버들이 갖고 있는 다양한 색상을 입힌다는 의미를 담고 있다.
현재 엔티티: 크레용팝
새 엔티티: 걸스데이
→ 수정된 문장: 걸스데이는 대한민국의 오인조 걸 그룹이다. '걸스데이'라는 이름은 음악이라는 도화지에 멤버들이 갖고 있는 다양한 색상을 입힌다는 의미를 담고 있다.

–
현재 엔티티: {answer}
새 엔티티: {new_entity}

문장:
{text_to_tts}

아래 형식으로만 출력하세요.
수정된 문장:

Stage 4: TTS Text Normalization

Objective: Convert numerals and non-Korean characters into spoken Korean for TTS synthesis.

아래 문장을 한국어 전사로 바꿔 주세요. 결과 문장은 한글만 포함해야 합니다.
[예시]
context:
황산벌 전투는 660년 8월 20일에서 8월 21일 황산벌에서 신라 군과 백제 군 사이에 일어났던 전투이다.
변환된 문장 : 황산벌 전투는 육백육십년 팔월 이십일에서 팔월 이십일일 황산벌에서 신라 군과 백제 군 사이에 일어났던 전투이다.

아래 문장을 한국어 전사로 바꿔 주세요. 결과 문장은 한글만 포함해야 합니다.
context:
{content}

변환된 문장:

Stage 5: Distractor Generation

Objective: Generate two additional distractors to form four-way multiple-choice options.

4지선다 문제를 만들기 위해 선택지 4개가 필요합니다. 정답은 이미 정해져 있습니다.

정답 (new_answer): {new_answer}
원래 답 (answer, 오답으로 사용): {answer}

위와 **같은 유형**의 비슷한 오답 2개를 더 만들어 주세요. (정답/원래 답과 같은 카테고리이지만 다른 항목)
예: 전투명 → 다른 전투명, 걸그룹명 → 다른 걸그룹명, 기관명 → 다른 기관명, 야구 선수 → 다른 야구 선수

아래 형식으로만 출력하세요.
오답1: (한 단어 또는 짧은 구)
오답2: (한 단어 또는 짧은 구)

Stage 6: Answer Verification

Objective: Verify that the generated question is answerable from the transformed context.

아래 질문에 대한 답을 컨텍스트에서 찾아서 답하세요. 알고 있던 내용과 다르더라도 컨텍스트에 있는 내용을 기반으로 답하세요.
[예시]
컨텍스트: 안시성 전투는 제2차 고수 전쟁을 고구려의 승리로 장식한 전투이다. 평양을 직공했던 수나라의 30만 군대가 살수를 건너 회군하던 중 을지문덕 군의 공격을 받아 궤멸적인 피해를 입고 패주했다. 귀주 대첩, 한산도 대첩과 함께 한국사 3대 대첩 중 하나로 불린다.
질문: 제2차 고수 전쟁에서 고구려의 승리를 이끈 전투로, 수나라의 30만 군대가 을지문덕 군의 공격을 받아 큰 피해를 입고 패주한 전투는 무엇인가?
보기: (1) 백강 전투 (2) 살수 대첩 (3) 황산 전투 (4) 안시성 전투
답: (4) 안시성 전투

아래 질문에 대한 답을 컨텍스트에서 찾아서 답하세요. 알고 있던 내용과 다르더라도 컨텍스트에 있는 내용을 기반으로 답하세요.
컨텍스트: {context}
질문: {question}
보기: {choices}
답:

A.11 EAR Qualitative Analysis Example

EAR Example Instance

PA-QA sample `000008` *used in Figure 3.*

Context:
저스틴이라는 이름의 어린 소년이 농장에서 살았습니다. 그의 아버지 제임스가 농장주였습니다. 그의 삼촌인 제리, 토드, 빌리도 농장에서 함께 일했습니다. 농장에는 닭, 소, 말, 양, 돼지가 있었습니다. 저스틴은 농장의 모든 동물들과 함께 노는 것을 좋아했지만, 특히 말과 함께 있는 것을 가장 좋아했습니다. 그는 하루 종일 말을 타며 시간을 보냈습니다. 특히 아기 말이 태어나는 것을 좋아했습니다. 아기 말은 봄에 태어났습니다. 아기 말이 태어나면 저스틴의 아버지는 아기 말을 외양간에서 데리고 나와 저스틴이 만지게 해주었습니다. 저스틴은 농장의 들판에서 노는 것도 좋아했습니다. 그는 들판을 뛰어다니며 나비를 잡으려고 했습니다. 또한 밤늦게까지 깨어 어둠 속에서 반딧불이를 잡는 것도 좋아했습니다. 저스틴은 여름에 낚시를 가는 것도 좋아했습니다. 어느 화요일 저녁, 그는 농장의 연못으로 가서 아주 큰 메기를 잡았습니다. 메기가 너무 커서 거의 낚싯대를 부러뜨릴 뻔했습니다. 저스틴은 메기를 집으로 가져왔습니다. 메기는 슬픈 눈으로 저스틴을 바라보았습니다. 메기를 바라보며 저스틴은 메기를 잡은 것이 미안했습니다. 그는 메기를 다시 연못에 던져주었습니다. 저스틴은 낚싯대를 정리하고 집으로 돌아갔습니다. 그는 아버지에게 메기를 잡았지만 다시 풀어주었다고 말했습니다. 아버지는 "네가 그런 일을 해서 자랑스럽다."라고 말했습니다.

Prompt_ko: 저스틴의 아버지는 왜 그를 자랑스러워했을까요?

Choices_ko:
(A) 그는 나비를 잡았다. (B) 그는 메기를 풀어주었다. (C) 그는 반딧불이를 쫓아다녔다. (D) 그는 말을 탔다.

Ground truth: (B)

Evidence: 그는 메기를 다시 연못에 던져주었습니다.

Non-evidence: 저스틴은 여름에 낚시를 가는 것도 좋아했습니다.

A.12 Human annotation

Item	Description
Task	TTS naturalness evaluation
Annotators	7 AI-major undergraduate/graduate students
Recruitment	Voluntary participants from our research lab
Compensation	No additional compensation provided
Consent	All annotators were informed of data usage and provided consent
Instruction	
TTS 합성이 자연스럽다면 1점, 그렇지 않고 부자연스럽거나 내용이 어색하다면 0점을 작성해주세요. *(Score 1 if the synthesized speech sounds natural; score 0 if it sounds unnatural or the content is awkward.)*	
Score 1	Natural-sounding synthesized speech
Score 0	Unnatural speech or awkward content

Table 9: Human annotation guidelines for TTS quality evaluation.