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/Koala-Bench/.

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 character error rate (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 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), Gemini-flash (Comanici et al., 2025), and Raon-speech (KRAFTON AI, 2026) with varying parameter scales, providing a comprehensive analysis of current LALMs on Korean speech understanding. Our main contributions are summarized as follows:

- 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.

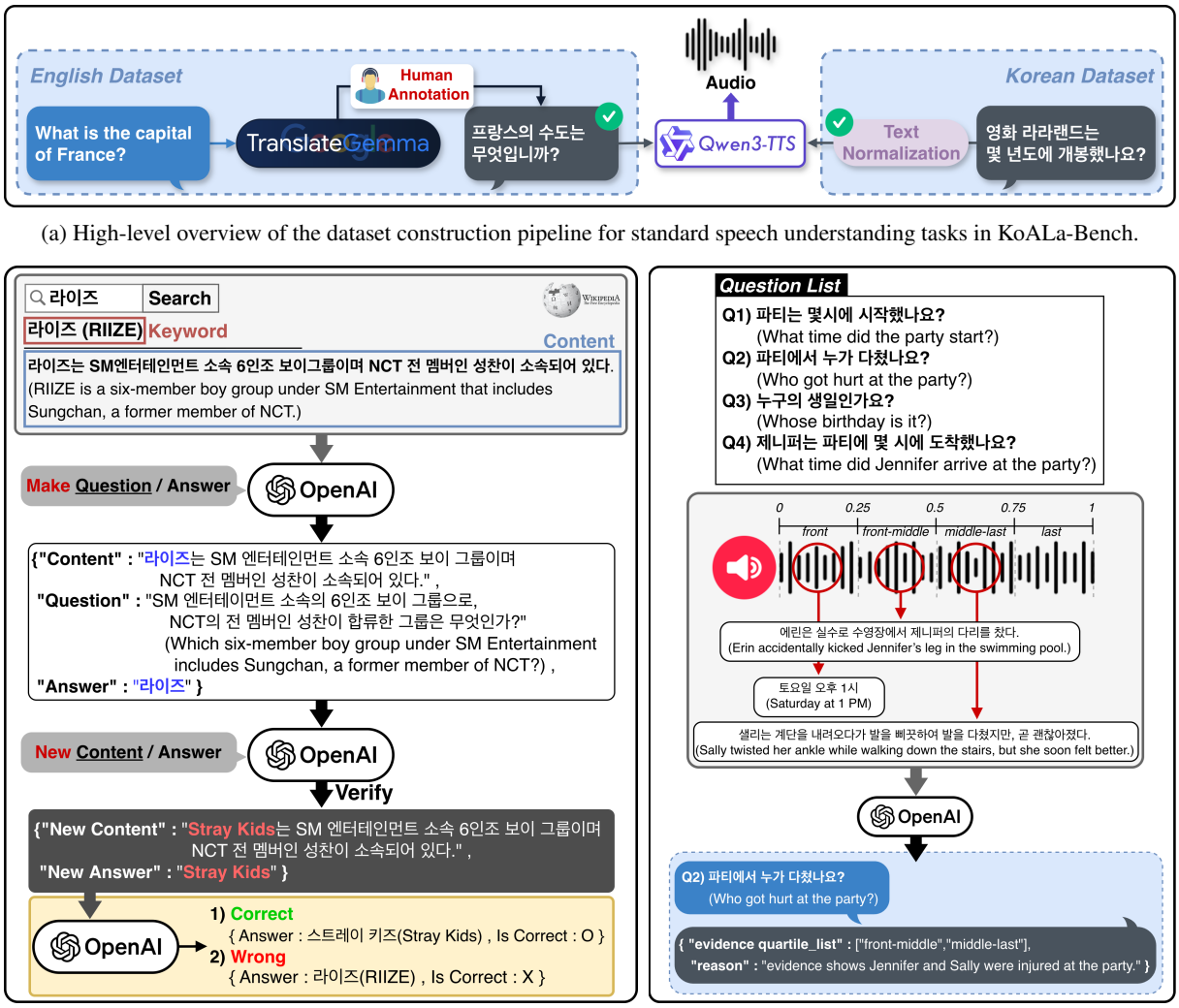
2 Related Works

Audio Benchmarks. 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) and AudioBench (Wang et al., 2025) evaluate audio understanding and instruction following across diverse audio inputs. Dynamic-SUPERB (Huang et al., 2024) focuses on diverse speech processing tasks including content, speaker, and semantic understanding, while AU-Harness (Surapaneni et al., 2025) focuses on temporal audio understanding and complex reasoning capabilities. More recent benchmarks explore specialized abilities, including audio dialogue understanding and expert-level reasoning (Gao et al., 2025; Sakshi et al., 2025). However, 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 a dedicated benchmark for evaluating Korean speech understanding capabilities of LALMs.

Faithfulness of Multimodal Language Models. Recent 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

We introduce KoALa-Bench, a benchmark designed to assess the speech understanding and speech faithfulness of Korean LALMs across six tasks, including ASR, ST, SQA, SIF, SCA-QA, and PA-QA. Following recent LALM benchmarks (Chen et al., 2026; Wang et al., 2025; Yan et al., 2025; Gao et al., 2025), KoALa-Bench includes synthesized speech in several subsets. To ensure



(a) High-level overview of the dataset construction pipeline for 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.

data integrity, we conduct quality verification by first removing samples with a CER exceeding 0.1, and then performing human annotation to further filter out unnatural samples. The instructions for human annotation are provided in Appendix A.2. An overview of the data construction pipeline is illustrated in Figure 1, and the detailed statistics of each subset are summarized in Table 1. The prompts for all tasks are provided in Appendix D.0.2. Below, we describe each subset along with its corresponding evaluation methodology.

3.1 Automatic Speech Recognition (ASR)

The ASR task is the most fundamental speech understanding task, evaluating a model's ability to transcribe spoken utterances into text. The model receives a target speech input along with a text prompt instructing it to transcribe the speech into Korean. The generated transcription is then compared with the ground-truth reference.

Datasets. We use three datasets for ASR evaluation: KsponSpeech (Bang et al., 2020), Common Voice Korean (Ardila et al., 2020), and Zeroth-Korean[1]. KsponSpeech provides spontaneous Korean conversational speech covering diverse everyday topics. Common Voice Korean and Zeroth-Korean contain speech recorded under diverse acoustic conditions as well as relatively clean read

[1] https://openslr.org/40/

speech, enabling evaluation across different speaking styles and recording environments.

Evaluation. 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)

The ST task evaluates cross-lingual speech translation from English speech into Korean text. We select English as the source language as it is the most frequently encountered foreign language for Korean users, making this the most practically relevant translation direction. The model receives an English speech input along with a text prompt instructing it to translate the speech into Korean.

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.

Evaluation. 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. For BERTScore, we use `xlm-roberta-large` (Conneau et al., 2020) as the backbone model.

3.3 Speech Question Answering (SQA)

The SQA task evaluates the model's ability to comprehend speech content and answer questions about it. The model receives a speech input as context, paired with a textual question and answer choices, and is required to select the correct option.

Datasets. For the SQA task, we evaluate on both short-form and long-form speech datasets. For short-form evaluation, we use two Korean QA benchmarks: CLIcK (Kim et al., 2024a) and KoBEST-BoolQ (Kim et al., 2022). CLIcK is a multiple-choice QA benchmark sourced from official exams and textbooks, covering 11 subcategories across language and culture domains. KoBEST-BoolQ is the Boolean QA subset of KoBEST, where each instance consists of a Korean Wikipedia passage and a yes/no question. Since both benchmarks are originally text-based, we synthesize the speech inputs using Qwen3-TTS (Hu et al., 2026) with additional quality verification.

For long-form evaluation, we construct long speech question and answering (LSQA) from the listening section of the KCSAT, a nationwide university entrance examination known for its high difficulty. We collect the original audio recordings from EBSi[4], which consist 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. A sample instance from the KCSAT dataset is provided in Appendix A.3.1.

Evaluation. For all SQA datasets, we use accuracy as the evaluation metric, measuring the proportion of correctly answered questions. We adopt two scoring approaches to determine the model's prediction: logit-based and generation-based. The logit-based approach computes the log-probability of each answer choice and treats the one with the highest value as the model's prediction. The generation-based approach extracts the predicted choice from the generated response via string matching.

3.4 Speech Instruction Following (SIF)

The SIF task evaluates the model's ability to understand and follow instructions delivered entirely in speech form. Unlike SQA, the model receives only a speech input containing the instruction, without any accompanying text, and is required to generate an appropriate response.

Datasets. We translate existing English instruction datasets into Korean using TranslateGemma-27B-IT (Finkelstein et al., 2026). The source datasets include ALPACA (Taori et al., 2023), Vicuna (Chiang et al., 2023), and OpenHermes[5]. We additionally include Korean-native instructions from KUDGE (Son et al., 2024). All instructions are then synthesized into speech using Qwen3-TTS (Hu et al., 2026) with additional quality verification.

[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

Evaluation. SIF performance is evaluated using GPT-4o as a judge model, following prior work on instruction following evaluation (Kim et al., 2025, 2024b; Wang et al., 2025). The judge model assesses whether the generated response correctly follows the instruction. Detailed evaluation prompts are provided in Appendix D.0.1.

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

Motivated by our observation, as illustrated in Appendix C.1, that models often ignore speech context, we introduce SCA-QA. Inspired by DisentQA (Neeman et al., 2023), the task embeds conflicting facts into the speech context and poses questions about them, evaluating whether the model grounds its response in the given acoustic context or relies on parametric knowledge.

Datasets. To construct the dataset, we generate 200 question-answer pairs for each topic using the GPT API, covering K-pop, K-history, and K-sports. We then replace key answer entities in both the question and the speech context with different entities (e.g., replacing "BTS" with "BLACKPINK"), so that the correct answer based on the speech context conflicts with world knowledge. This allows us to evaluate whether the model faithfully grounds its answer in the given speech context rather than relying on its parametric knowledge. The construction pipeline and prompts are provided in Appendix D.0.3. All samples were filtered through human annotation, where annotators assessed the appropriateness of each QA pair and the validity of the conflicting facts. The model receives a speech context containing the conflicting fact, along with a textual question and four answer choices. A detailed example is provided in Appendix A.3.2.

Evaluation. 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 B.

3.6 Postion Aware Question Answering (PA-QA)

Existing speech benchmarks often rely on short utterances and lack evidence-position annotations for each QA pair (Chen et al., 2026; Wang et al., 2025), making it difficult to analyze which parts of a speech input a model effectively uses or fails to use when answering a question. To address this limitation, KoALa-Bench introduces PA-QA, a position-aware evaluation setup designed to examine model performance according to the location of answer-relevant evidence within long speech inputs.

Datasets. We construct the dataset from MCTest (Richardson et al., 2013), a machine comprehension dataset of short stories and multiple-choice questions. For each QA pair, we first split the speech context into individual sentences, treating them as possible evidence candidates. We then use the question and answer to select the sentences that support the answer, with GPT-4o-mini used for candidate selection and verification. Each verified evidence sentence is mapped to its relative position in the speech context, which is normalized to [0,1] and divided into four temporal segments: front (0–0.25), front-middle (0.25–0.5), middle-late (0.5–0.75), and late (0.75–1.0). This setup allows us to analyze whether model performance changes depending on where the relevant information appears in long speech inputs. The full evidence grounding procedure is described in Appendix A.5. Speech contexts are synthesized from textual passages using Qwen3-TTS (Hu et al., 2026), and supporting evidence sentences are identified with GPT-4o-mini. Also, all samples were verified through the quality control process. A sample instance is provided in Appendix A.3.3.

Evaluation. We adopt two evaluation methods for PA-QA. The first follows the same protocol as in SQA and report accuracy over all QA pairs, scoring whether the model selects the correct answer. For position-aware analysis, we group QA pairs by the temporal segment containing their supporting evidence and report segment-wise accuracy, allowing us to examine whether models are more vulnerable to relevant information appearing at specific positions in the speech input.

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	4.03	5.73	0.32	39.62
ETRI-TST-HE			✗	544	1.07	7.09	0.77	40.79
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.

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], Gemini-flash-lite (Comanici et al., 2025) and Raon-Speech-9B (KRAFTON AI, 2026). 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-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.1.

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

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

Since LALMs can be highly sensitive to prompt phrasing (Maia Polo et al., 2024; Sclar et al., 2024; Zhuo et al., 2024), we evaluate each task using four Korean prompts with slight lexical and structural variations and report the average score as the main result. The full prompt list, including best scores for diagnostic purposes, is provided in Appendix D.0.2.

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. In ASR, Qwen3 and Raon consistently outperformed other models, with Qwen3 achieving the lowest CER on read speech (Zeroth) and Raon excelling on spontaneous and conversational speech (Common Voice, KsponSpeech). Qwen3 also achieved the highest BERTScore on ST, while Gemini-flash led on short-form SQA benchmarks and Qwen3 on the long-form KCSAT. For SIF, GPT achieved the highest scores across all datasets, indicating stronger instruction-following capabilities. Under noisy conditions, ASR tasks showed noticeable CER increases, whereas SQA and SIF exhibited modest degradation, suggesting that back-

Dataset		Metric	Qwen3-omni	Gemma-3n	GPT-audio	Voxtral	Gemini-flash	Raon
ASR								
Zeroth		CER ↓	**3.39** / **4.02**	100↑ / 100↑	6.87 / 9.00	46.29 / 42.91	13.60 / 14.56	4.28 / 5.27
Common Voice			4.99 / **7.02**	100↑ / 100↑	33.05 / 36.21	65.80 / 63.57	13.74 / 26.74	**4.78** / 7.39
KsponSpeech-Clean			8.58	100↑	100↑	67.83	83.19	**8.16**
KsponSpeech-Other			8.15	100↑	63.64	63.15	45.14	**7.81**
ST								
ETRI-TST-COMMON		BERTScore ↑	**93.38**	85.02	93.10	92.64	91.60	92.67
ETRI-TST-HE			**93.89**	85.82	93.69	92.98	92.17	93.22
SQA								
Short-Form	CLIcK	Accuracy ↑	63.57 / 62.01	35.11 / 35.31	61.64 / 60.07	41.94 / 41.88	**67.27** / **66.69**	55.68 / 55.18
	KoBEST BoolQ		50.40 / 49.84	50.36 / 50.27	51.88 / 50.45	50.49 / 50.40	**52.86** / **54.92**	50.45 / 50.51
Long-Form	KCSAT		80.59 / **82.36**	33.53 / 37.65	52.90 / 47.10	69.41 / 71.76	**81.18** / 78.82	45.88 / 45.29
SIF								
KUDGE		Score (GPT as Judge) ↑	70.89 / 70.50	70.68 / 70.53	**74.07** / **73.99**	59.14 / 59.07	70.29 / 70.39	68.35 / 66.85
Vicuna			76.50 / 71.79	79.38 / 78.84	**82.14** / **81.79**	62.82 / 64.86	76.43 / 73.29	73.81 / 71.52
OpenHermes			84.09 / 80.74	82.93 / 84.25	**89.42** / **89.62**	65.82 / 64.87	82.69 / 80.71	78.65 / 78.30
Alpaca			82.55 / 80.68	82.83 / 82.20	**90.58** / **90.58**	68.48 / 68.16	85.94 / 86.88	76.34 / 75.80

Table 2: Comparison across ASR, ST, SQA, and SIF. All results are averaged over four prompts using greedy decoding. 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	Raon
SCA-QA								
K-history (Before)		Text-only Accuracy ↑	56.44	46.78	69.30	38.37	**79.21**	50.25
K-history (After Chosun)			77.74	67.38	79.30	47.26	**87.80**	61.59
K-sports			52.56	55.68	69.30	30.40	**78.41**	48.58
K-pop			59.47	61.17	69.90	36.89	**89.32**	51.70
K-history (Before)		SCF Score ↑	**93.86**	49.72	61.40	77.24	66.25	36.91
K-history (After Chosun)			**85.46**	36.10	32.30	76.57	59.72	21.27
K-sports			84.34	69.33	39.30	**88.85**	86.96	21.89
K-pop			**90.60**	61.88	37.50	87.39	78.26	25.10
PA-QA								
MC-Test Total		Accuracy ↑	**92.51** / **93.12**	48.70 / 48.78	77.23 / 79.69	84.25 / 85.24	92.31 / 92.00	40.98 / 41.28
Position - Aware	Front		91.55 / 92.07	46.72 / 46.55	72.05 / 75.78	84.31 / 85.34	**95.03** / **93.17**	43.48 / 42.03
	Front-Middle		**93.15** / **94.35**	50.30 / 50.00	83.52 / 83.52	88.10 / 88.10	89.01 / 87.91	42.70 / 44.94
	Middle-Last		**90.26** / **89.29**	48.38 / 50.00	74.39 / 70.73	82.79 / 84.42	82.93 / 85.37	32.47 / 32.47
	Last		**95.31** / 95.83	44.79 / 43.75	80.39 / 86.27	83.85 / 78.65	92.16 / **98.04**	40.00 / 38.18

Table 3: Performance comparison on speech faithfulness tasks. All results are averaged over four prompts using greedy decoding. 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.

ground noise affects low-level recognition more than higher-level understanding.

4.2.2 Speech Faithfulness

SCA-QA. As shown in Table 3, SCA-QA highlights a gap between Korea-specific parametric knowledge and speech-grounded faithfulness. Gemini-flash achieved the highest Text-only accuracy across all domains, and GPT-audio also showed strong Text-only performance; however, both obtained much lower SCF scores than Qwen3-Omni and Voxtral, indicating reliance on prior knowledge rather than counterfactual speech evidence. In contrast, Qwen3-Omni achieved the highest SCF score in three of four domains, while Voxtral showed strong faithfulness on K-sports despite lower Text-only accuracy. Raon performed poorly in both Text-only accuracy and SCF, suggesting limited ability to identify the answer from counterfactual speech input. Additional analysis is provided in Section 5.

PA-QA. On PA-QA (Table3), we report overall MC-Test and position-aware accuracy to examine evidence grounding across speech positions. Qwen3-Omni achieved the highest overall accuracy under both clean and noisy conditions, with the best performance in Front-Middle and Middle-Last, while Gemini-flash was strongest in Front and noisy Last. These results suggest that Qwen3-

	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.

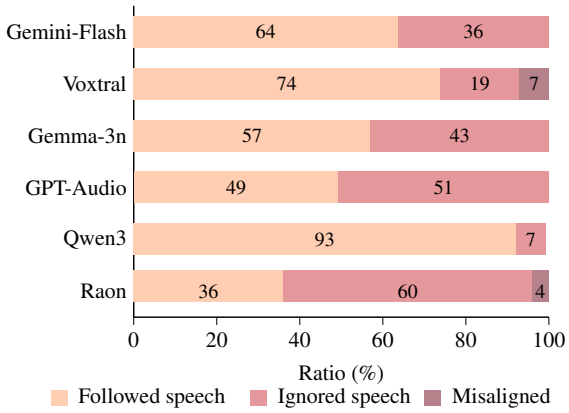


Figure 2: Model behavior under correct speech contexts on the SCA-QA benchmark. The bar chart reports results for the K-history (before Chosun) subset.

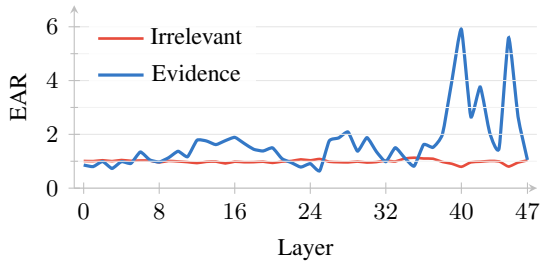


Figure 3: Per-layer EAR of the evidence sentence vs. irrelevant sentences (details of the sample are provided in Appendix C.4). The model allocates substantially more attention to the evidence in deeper layers (L39–L45).

Omni provides more consistent grounding across positions, whereas Gemini-flash is particularly effective for early or noisy late evidence. Gemma-3n and Raon remained substantially weaker, indicating limited long-form speech understanding and evidence grounding. Additional analysis is provided in Section 5.

5 Analysis

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 Voxtral and Raon exhibit speech–text alignment failures, with Raon showing a strong tendency to ignore the speech context. This demonstrates that SCA-QA enables fine-grained evaluation of speech faithfulness, extending to speech–text alignment analysis.

PA-QA. To analyze whether models attend to relevant speech evidence, we computed the evidence attention ratio (EAR) (Appendix B.3) using teacher-forced generation, where the model-generated answer tokens are appended to the prompt and processed in a single forward pass to extract attention weights. We measured the attention from the answer token to each speech sentence. As shown in Figure 3, EAR for the evidence sentence (located at the 47% position of the story) exceeded 1.0 in most layers, with strong peaks in layers 39-45, while irrelevant sentences remained near 1.0 across all layers. Overall, the evidence sentence showed a substantially higher average EAR than irrelevant sentences (1.68 vs. 0.99), suggesting that the model selectively attends to the ground-truth evidence even when it appears in the middle of long speech input. Notably, this analysis is enabled by the explicit evidence annotations in PA-QA, which allow fine-grained examination of whether a model's attention aligns with the intended evidence positions.

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. We use the GPT and Gemini APIs for evaluation, and the GPT API is additionally used for dataset construction. 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.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. 2026. Voicebench: Benchmarking llm-based voice assistants. *Transactions of the Association for Computational Linguistics*, 14:378–398.

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*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

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.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo,

Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. Translategemma technical report. *Preprint*, arXiv:2601.09012.

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*.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, and 1 others. 2024. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12136–12140. IEEE.

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*.

Dohyeong Kim, Myeongjun Jang, Deuk Sin Kwon, and Eric Davis. 2022. Kobest: Korean balanced evaluation of significant tasks. *arXiv preprint arXiv:2204.04541*.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024a. Click: A benchmark dataset of cultural and linguistic intelligence in korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346.

Seungone Kim, Jay Shin, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Ryan Shin, Sungdong Kim, James Thorne, Minjoon Seo, and 1 others. 2024b. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, volume 2024, pages 29927–29962.

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, and 1 others. 2025. The biggen bench: A principled benchmark for fine-grained evaluation of language models with 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 5877–5919.

KRAFTON AI. 2026. Raon-speech technical report.

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*.

Sidharth Surapaneni, Hoang Nguyen, Jash Mehta, Aman Tiwari, Oluwanifemi Bamgbose, Akshay Kalkunte, Sai Rajeswar, and Sathwik Tejaswi Madhusudhan. 2025. Au-harness: An open-source toolkit for holistic evaluation of audio llms. *arXiv preprint arXiv:2509.08031*.

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.

Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. 2025. Uro-bench: Towards comprehensive evaluation for end-to-end spoken dialogue models. *arXiv preprint arXiv:2502.17810*.

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.

Appendices

A Data Construction Details

This section details the dataset construction and evaluation setup of KoALa-Bench. We describe the noise augmentation protocol, human annotation guidelines, representative examples from each evaluation subset, the SCA-QA construction pipeline, and the two-stage evidence grounding procedure for PA-QA.

A.1 Details of Noise Injection for Robustness Experiments

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

Table 5: 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[8] 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 5. 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.2 Human annotation

We conduct a binary human evaluation to filter out unnatural TTS samples. The synthesized dataset is evenly divided among seven annotators, all native Korean speakers majoring in AI, where each annotator is responsible for reviewing their assigned partition. Each sample is rated as either natural (1) or unnatural/awkward (0) following the guidelines in Table 6. Samples rated as unnatural are discarded from the final dataset. All annotators participated voluntarily as part of a collaborative research effort.

[8] https://selectstar.ai/

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 6: Human annotation guidelines for TTS quality evaluation.

A.3 Dataset Examples

This section presents representative examples from the three evaluation datasets in KoALa-Bench: KCSAT, SCA-QA, and PA-QA. Section A.3.1 shows a KCSAT example consisting of a long-form dialogue-style speech context, a multiple-choice question, and its answer. Section A.3.2 provides SCA-QA examples from K-history, K-sports, and K-pop, illustrating how entity replacement modifies original speech contexts to construct counterfactual pairs; the keyword column denotes the replacement entity. Section A.3.3 presents a PA-QA example with evidence sentences mapped to their relative positions in the speech context, supporting fine-grained analysis of where answer-relevant information appears.

A.3.1 KCSAT

Table 7 presents a representative example from the KCSAT subset, which is derived from the listening section of the Korean College Scholastic Ability Test (KCSAT).[9] We use all listening examinations administered between 2006 and 2012, after which the Korean listening section was discontinued.Each instance is constructed at the question level by segmenting the listening audio into a long-form speech context paired with a multiple-choice question and its ground-truth answer. The questions cover diverse discourse types, including lectures, presentations, debates, interviews, and multi-turn dialogues,

[9] Official materials are available at https://www.ebsi.co.kr/

Field	Content
Speech Context	**진행자** : 안녕하세요? 전문가와 함께 하는 삶의 이야기 시간입니다. 오늘은 건축가이신 김 선생님을 모시고 이야기를 들어 볼까 합니다. 안녕하세요? 선생님. *(**Host**: Hello? Welcome to Life Stories with Experts. Today, we have architect Mr. Kim with us to share his thoughts. Welcome, Mr. Kim.)* **김 선생님** : 네, 안녕하세요? 오늘은 제가 그동안 건축 분야에서 일하면서 느꼈던 계단과 우리네 삶의 의미에 대해 이야기를 해 보겠습니다. 누워서 세상을 보던 아이가 기기 시작하면서 보는 세상은, 이전과는 다른 새로운 세계입니다. 아이가 첫발을 내디디며 느끼는 것도 달라진 높이에서 보는 기쁨이죠. 일상적인 삶의 공간에 위치하고 있는 계단도 이런 즐거움과 삶의 의미를 전해 주죠. *(**Mr. Kim**: Yes, hello. Today, I would like to talk about stairs and the meaning they hold in our lives, based on my experience working in architecture. When a baby who once saw the world lying down begins to crawl, the world becomes an entirely new place. The joy a child feels when taking their first steps also comes from seeing things at a different height. Stairs, situated in our everyday living spaces, convey this same kind of joy and meaning in life.)* **진행자** : 네. 그렇다면 계단이 우리 삶에서 갖는 의미는 어떤 것인가요? *(**Host**: I see. Then, what meaning do stairs hold in our lives?)* **김 선생님** : 아래층과 위층은 들어오는 빛, 경치, 심지어는 냄새까지도 다른 공간입니다. 계단은 이렇게 분리된 두 공간을 이어 주고 있습니다. 계단을 오르내리면서 서로 다른 차원의 공간과 연결될 수 있는 거지요. 이렇게 우리는 계단을 통해 새로운 세계와 교류하게 되고, 달라진 높이에서 세상을 새롭게 조망할 수 있게 되는 것이죠. 결국 인간은 계단을 통해 *(**Mr. Kim**: The lower and upper floors are different spaces in terms of light, scenery, and even smell. Stairs connect these two separate spaces. By going up and down the stairs, we can be connected to spaces of different dimensions. In this way, through stairs, we come to interact with a new world and gain a fresh perspective on the world from a different height. Ultimately, through stairs, humans...)*
Question	방송의 마지막에 이어질 말로 가장 적절한 것은? *(What is the most appropriate statement to follow at the end of the broadcast?)*
Choices	(1) 세상에 대한 긍정적인 시선을 갖게 됩니다. *(We come to have a positive outlook on the world.)* (2) 어려운 이웃을 배려하는 마음을 배우게 됩니다. *(We learn to care for neighbors in need.)* (3) 새로운 세상과 만나고 소통하는 기쁨을 얻게 됩니다. *(We gain the joy of encountering and communicating with a new world.)* (4) 끊임없는 도전만이 성공을 위한 열쇠라는 것을 알게 됩니다. *(We learn that only constant challenge is the key to success.)* (5) 일의 성패는 스스로의 노력에 달려 있다는 것을 배우게 됩니다. *(We learn that the success or failure of a task depends on one's own effort.)*
Answer	(3)

Table 7: A representative example from the KCSAT subset. The example requires selecting the most appropriate continuation by reasoning over the dialogue context.

and require various reasoning skills such as identifying the speaker's intent, inferring the most appropriate continuation of an incomplete utterance, and comprehending the main argument or factual details from the speech context. In the example shown in Table 7, the final utterance is incomplete, and the model must select the most appropriate continuation from five candidates. The correct answer is (3), since the preceding dialogue states that stairs connect separate spaces, allow people to interact with a new world, and provide a new perspective from a different height. Thus, the example requires discourse-level inference over the dialogue context rather than matching a single explicit phrase in the speech.

A.3.2 SCA-QA

We collect 200 samples for each of four domains: pre-Joseon Korean history, post-Joseon Korean history, K-sports, and K-pop, yielding 800 initial samples. After human validation for counterfactual consistency, 374 samples are retained in the final SCA-QA dataset. Tables 8, 9, and 10 show representative counterfactual examples from K-history, K-sports, and K-pop, respectively. Each example pairs an original speech context with a counterfactual version created by replacing the target entity while preserving the question structure. The keyword field specifies the entity that determines the speech-grounded answer.

K-history. Table 8 shows a K-history example where the target entity is replaced from Sejong the Great to Yi Sun-sin, changing the answer from the

	Original	Counterfactual
Speech Context	세종대왕 동상은 대한민국 서울특별시 종로구 세종로 광화문 광장에 있는 세종대왕 동상이다. 세종대왕 동상은 홍익대학교 교수이자 조각가인 김영원이 설계한 것으로, 2009년 건립되었다. *(Sejong the Great Statue is a statue of Sejong the Great located in Gwanghwamun Square, Sejong-ro, Jongno-gu, Seoul, Republic of Korea. The statue was designed by sculptor Kim Young-won, a professor at Hongik University, and was erected in 2009.)*	이순신 동상은 대한민국 서울특별시 종로구 세종로 광화문 광장에 있는 동상이다. 이순신 동상은 홍익대학교 교수이자 조각가인 김영원이 설계한 것으로, 2009년 건립되었다. *(Yi Sun-sin Statue is a statue of Yi Sun-sin located in Gwanghwamun Square, Sejong-ro, Jongno-gu, Seoul, Republic of Korea. The statue was designed by sculptor Kim Young-won, a professor at Hongik University, and was erected in 2009.)*
Question	대한민국 서울특별시 종로구 세종로 광화문 광장에 위치하며, 홍익대학교 교수인 김영원이 설계하고 2009년에 건립된 동상의 이름은 무엇인가? *(What is the name of the statue located at Gwanghwamun Square in Sejong-ro, Jongno-gu, Seoul, Republic of Korea, which was designed by Kim Young-won, a professor at Hongik University, and erected in 2009?)*	
Keyword	세종대왕 *(Sejong the Great)*	이순신 *(Yi Sun-sin)*
Answer	세종대왕 동상 *(Statue of Sejong the Great)*	이순신 동상 *(Statue of Yi Sun-sin)*

Table 8: A representative K-history example from SCA-QA. The counterfactual context replaces the target entity from Sejong the Great to Yi Sun-sin, changing the speech grounded answer from the Sejong the Great Statue to the Yi Sun-sin Statue.

	Original	Counterfactual
Speech Context	유창혁은 대한민국의 바둑 기사이다. 1966년에 서울특별시에서 태어났고, 1979년에 어린이 국수전 3연패 학초배 우승하였다. 1984년 입단 이후 신인 대회 수상을 시작으로 세계대회 우승 6회 등 세계대회 그랜드슬램을 달성한 바둑 기사이다. *(Yoo Chang-hyuk is a Korean professional Go player. He was born in Seoul in 1966 and won three consecutive Children's Guksu titles and the Hakcho Cup in 1979. After turning professional in 1984, he began winning rookie tournaments and went on to achieve a Grand Slam in international competitions, with six world championship titles.)*	이세돌은 대한민국의 바둑 기사이다. 1966년에 서울특별시에서 태어났고, 1979년에 어린이 국수전 3연패 학초배 우승하였다. 1984년 입단 이후 신인 대회 수상을 시작으로 세계대회 우승 6회 등 세계대회 그랜드슬램을 달성한 바둑 기사이다. *(Lee Sedol is a Korean professional Go player. He was born in Seoul in 1966 and won three consecutive Children's Guksu titles and the Hakcho Cup in 1979. After turning professional in 1984, he began winning rookie tournaments and went on to achieve a Grand Slam in international competitions, with six world championship titles.)*
Question	대한민국의 바둑 기사로, 1966년에 서울특별시에서 태어나 어린이 국수전 3연패와 학초배 우승을 차지한 후, 1984년 입단하여 세계대회에서 6회 우승을 기록한 인물은 누구인가? *(Who is the Korean professional Go player, born in Seoul in 1966, who won three consecutive Children's Guksu titles and the Hakcho Cup, turned professional in 1984, and went on to win six world championship titles?)*	
Keyword	유창혁 *(Yoo Chang-hyuk)*	이세돌 *(Lee Sedol)*
Answer	유창혁 *(Yoo Chang-hyuk)*	이세돌 *(Lee Sedol)*

Table 9: A representative K-sports example from SCA-QA. The counterfactual context replaces the target entity from Yoo Chang-hyuk to Lee Sedol, requiring the answer to change according to the entity stated in the speech context.

Sejong the Great Statue to the Yi Sun-sin Statue according to the speech context.

K-sports. Table 9 presents a K-sports example involving Korean professional Go players. The original context identifies Yoo Chang-hyuk, while the counterfactual context replaces him with Lee Sedol while preserving the surrounding biographical description. Thus, the model must update its answer based on the speech context rather than relying on prior knowledge about the described attributes.

K-pop. Table 10 shows a K-pop example where the described music group is changed from Davichi to MAMAMOO. Although the debut date and album information remain unchanged, the correct answer must follow the group name explicitly stated in the speech. This tests whether the model grounds

	Original	Counterfactual
Speech Context	다비치는 2008년 2월 22일 데뷔한 대한민국의 여성 보컬 듀오이다. 2008년, 정규 1집 Amaranth를 발매하며 성공적으로 데뷔했다. *(Davichi is a South Korean female vocal duo that debuted on February 22, 2008. They made a successful debut in 2008 with the release of their first studio album, Amaranth.)*	마마무는 2008년 2월 22일 데뷔한 대한민국의 여성 보컬 듀오이다. 2008년, 정규 1집 Amaranth를 발매하며 성공적으로 데뷔했다. *(MAMAMOO is a South Korean female vocal duo that debuted on February 22, 2008. They made a successful debut in 2008 with the release of their first studio album, Amaranth.)*
Question	2008년 2월 22일에 데뷔하여 정규 1집 Amaranth를 발매하며 성공적으로 출발한 대한민국의 여성 보컬 듀오는 무엇인가? *(What is the South Korean female vocal duo that debuted on February 22, 2008, and made a successful start with the release of their first studio album, Amaranth?)*	
Keyword	다비치 *(Davichi)*	마마무 *(MAMAMOO)*
Answer	다비치 *(Davichi)*	마마무 *(MAMAMOO)*

Table 10: A representative K-pop example from SCA-QA. The counterfactual context replaces the target entity from Davichi to MAMAMOO, changing the correct answer while preserving the overall question structure.

Field	Content
Speech Context	동물원 밖은 평소보다 따뜻했다. 광대는 분장을 한 채 땀을 흘리고 있었다. 하지만 그는 여전히 미소를 지으며 사람들을 웃게 만들었다. 그는 열심히 일했다. 누군가의 아들이 부탁해서 그는 종이 비행기를 던졌다. 작은 소녀가 그에게 "이렇게 하면 더 예뻐 보일 거예요"라고 말하자, 그는 심지어 얼굴에 젤리를 발랐다. 정오에는 접시를 던지며 저글링을 했고, 오후 1시에는 앉아서 점심을 먹었다. 요리사는 광대가 좋아하는 스타일로 샐러드를 만들어 주었고, 광대는 기분 좋게 샐러드를 먹었다. 동물원을 방문한 사람들이 그를 가리키며 미소를 짓는 것을 보면서 그는 유명해진 기분을 느꼈다. 그날은 더웠지만, 사람들의 반응을 보면서 광대는 결국 좋은 하루였다고 생각했다. 얼마 지나지 않아 그는 다시 저글링을 하고 농담을 하며 일에 몰두했다. 심지어 누군가가 그의 가짜 꽃 냄새를 맡자 물을 뿌리기도 했다. *(It was unusually warm outside the zoo, and the clown was sweating in his costume. Still, he kept smiling and worked hard to entertain visitors. At a child's request, he threw a paper airplane, and when a little girl suggested it, he even put jelly on his face. He later juggled plates, ate his favorite salad for lunch, and felt pleased as visitors pointed at him and smiled. Although the day was hot, he thought it had been a good day and soon returned to juggling, joking, and spraying water from his fake flower.)*
Question	광대는 어떻게 그렇게 열심히 일할까요? *(How does the clown work so hard?)*
Choices	(A) 그는 동물원에 머물렀다. *(He stayed at the zoo.)* (B) 그는 비행기 장난감을 던지고 얼굴에 젤리를 발랐다. *(He threw a toy airplane and put jelly on his face.)* (C) 그는 웃었다. *(He laughed.)* (D) 그는 밖에서 서 있었다. *(He stood outside.)*
Evidence	누군가의 아들이 부탁해서 그는 종이 비행기를 던졌다. 작은 소녀가 그에게 "이렇게 하면 더 예뻐 보일 거예요"라고 말하자, 그는 심지어 얼굴에 젤리를 발랐다. *(Someone's son asked, so he threw a paper airplane. When a little girl told him "You'll look prettier if you do this," he even put jelly on his face.)*
Position	Front, Front–middle
Answer	(B)

Table 11: A representative example from the PA-QA dataset. Each instance contains a long-form speech context, a multiple-choice question, the ground-truth answer, and evidence sentences annotated with their relative positions in the speech context.

its response in the provided speech context when it conflicts with potentially memorized associations.

Overall, these examples show that SCA-QA evaluates whether models ground their answers in the provided speech context rather than relying on parametric knowledge.

A.3.3 PA-QA

Table 11 shows an example from the PA-QA dataset, which evaluates position-aware question answering over long-form speech inputs. Each example consists of a speech context, a multiple-choice question, and the corresponding evidence

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 $\neq a'$;

sentences that support the correct answer. The evidence sentences are identified through the evidence grounding pipeline and mapped to their relative positions within the speech context, which is normalized and divided into four temporal segments: front (0–0.25), front-middle (0.25–0.5), middle-last (0.5–0.75), and last (0.75–1.0). In the example shown, the evidence spans the front and front-middle segments, indicating that the supporting information appears in the first half of the speech context. This example illustrates how PA-QA enables fine-grained analysis of whether LALMs can faithfully utilize information appearing at different locations within long-form speech inputs.

A.4 SCA-QA Pipeline

Algorithm 1 summarizes the construction procedure for SCA-QA. Given a Wikipedia passage x, we use the LLM client $\mathcal{G}$ with 3-shot prompting to generate a question–answer pair (q, a). We then generate a counterfactual entity a' that has the same semantic type as a but does not appear in x, and replace all occurrences of a with a' to obtain a counterfactual context x'. The question q is kept unchanged, so the correct answer becomes a' only when the model grounds its prediction in the modified context rather than relying on the original factual association with a. The counterfactual context is normalized into spoken Korean form and synthesized into speech using the TTS engine $\mathcal{T}$. We construct a four-way multiple-choice set $\mathcal{C}$ from the counterfactual answer a', the original answer a, and two same-type distractors generated by $\mathcal{G}$. To ensure sample validity, we retain only examples where $\mathcal{G}$ selects a' when given $(x', q, \mathcal{C})$, filtering out ambiguous or inconsistent cases.

A.5 Two Stage Evidence Grounding for PA-QA

Algorithm 2 summarizes our two-stage procedure for grounding answer-relevant evidence from text to audio. In the first stage, each sample is decomposed into its context x, question q, and gold answer $a^\star$. We split x into sentences and identify candidate evidence sentences using answer-based rule matching, with $\mathcal{G}$ used to resolve ambiguous cases or to select evidence directly when no rule-based candidate is found. The selected evidence set

Algorithm 2: Two-Stage Evidence Grounding: Text → Audio

Input: Samples $\mathcal{D}$, GPT client $\mathcal{G}$, ASR model $\mathcal{W}$
Output: Samples with evidence sentences, validity labels, text positions, 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` ← `False`; continue;
  Split x into sentences $\{s_i\}_{i=1}^N$ and record their text spans;
  Construct a candidate evidence pool $\mathcal{P}$ using answer-based rule matching;
  if $\mathcal{P} \neq \emptyset$ **then**
    Select initial evidence sentences $\mathcal{E}$ from $\mathcal{P}$, using $\mathcal{G}$ when candidate selection is ambiguous;
  else
    Use $\mathcal{G}$ to select evidence sentences from the sentence pool $\{s_i\}_{i=1}^N$;
  Verify $\mathcal{E}$ with $\mathcal{G}$ given $(q, a^\star, \mathcal{E})$;
  if *the verification indicates that additional evidence is needed* **then**
    Select an additional evidence sentence using $\mathcal{G}$ and update $\mathcal{E}$;
  Re-verify the final evidence set $\mathcal{E}$ with $\mathcal{G}$;
  Set `is_valid` ← (`supported` ∧ ¬`need_more`);
  Map each evidence sentence in $\mathcal{E}$ to its relative text position by normalizing its start position within x;
  Save evidence sentences, text spans, relative positions, temporal segment labels, verification result, and `is_valid`;
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* $s \in \mathcal{E}$ **do**
    Align s to the transcript using exact matching followed by fuzzy matching;
    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;

$\mathcal{E}$ is then verified by $\mathcal{G}$ with respect to $(q, a^\star)$, and additional evidence is added when the answer is not fully supported. We retain only samples whose final evidence set is verified as sufficient, and record each evidence sentence with its text span, normalized relative position, and temporal segment label. In the second stage, we map the verified text evidence to the audio signal. For each valid sample, the audio is transcribed using the ASR model $\mathcal{W}$ with word-level timestamps. Each evidence sentence in $\mathcal{E}$ is aligned to the transcript using exact matching followed by fuzzy matching, and the matched text span is converted into an audio time span $[t_s, t_e]$. This procedure produces samples with both text-level evidence positions and audio-level evidence spans, enabling position-aware analysis of whether models use the relevant part of the speech input.

B Evaluation Details

This section details the evaluation protocols used in KoALa-Bench. We describe the normalization pipeline for Korean ASR, formally define the SCF-score for measuring speech-context faithfulness in SCA-QA, and introduce the Evidence Attention Ratio (EAR) for analyzing whether models attend to answer-relevant evidence in long speech inputs.

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 12: 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%).

B.1 CER Normalization and Computation

Normalization Protocol. We evaluate Korean ASR using character error rate (CER) with a normalization pipeline tailored to the linguistic properties of Korean numerals. 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, as even native speakers may segment words inconsistently. Naïvely computing CER on such pairs would incur false penalties despite semantically identical transcriptions.

Normalization Pipeline. To address this, 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 12 shows concrete examples covering the normalization rules.

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.

B.2 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.

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 한 자 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	K/DA
Model Response *w/ question-only*	ITZY
Model Response *w/ speech + question*	ITZY

Table 13: A failure case where Qwen3-Omni-30B-A3B-Instruct does not leverage the provided speech context. Although the speech context states that the fandom name "Blaze" corresponds to K/DA, the model outputs the same incorrect answer, "ITZY," in both the question-only and speech-conditioned settings.

B.3 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:

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

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

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 employ teacher-forced generation: the model-generated answer tokens are appended to the prompt, and a single forward pass is performed with standard attention to capture the full attention weight matrix. EAR is then computed from the attention of the answer token (e.g., "(B)") to all speech tokens, averaged across attention heads. To compare evidence and non-evidence regions, we apply the same formula to each sentence individually, treating its tokens as E and all remaining speech tokens as N. We then report EAR for the evidence sentence and the average EAR across all non-evidence sentences.

C Additional Results and Analysis

This section provides supplementary analyses for KoALa-Bench, covering an EAR qualitative example, a speech-context failure case, additional speech translation metrics, and best prompt results across four prompt variants.

C.1 Failure to Leverage Speech Context

Table 13 shows a qualitative failure case where Qwen3-Omni-30B-A3B-Instruct does not leverage the provided speech context. Although the speech input explicitly states that the fandom name "Blaze" corresponds to K/DA, the model outputs the same incorrect answer, "ITZY," in both the question-only and speech-conditioned settings. This example illustrates that the model can rely on an unsupported prior association rather than grounding its prediction in the answer-relevant evidence contained in the speech input.

Dataset		Metric	Qwen3-omni	Gemma-3n	GPT-audio	Voxtral	Gemini-flash	Raon
ASR								
Zeroth		CER ↓	**3.33** / **3.91**	100↑ / 100↑	6.87 / 9.00	40.92 / 39.07	13.60 / 14.56	4.09 / 5.04
Common Voice			4.96 / **6.78**	100↑ / 100↑	33.05 / 36.21	60.10 / 58.98	13.74 / 26.74	**4.74** / 7.04
KsponSpeech-Clean			8.46	100↑	100↑	62.62	83.19	**7.55**
KsponSpeech-Other			7.91	100↑	63.64	56.04	45.14	**7.53**
ST								
ETRI-TST-COMMON		BERTScore ↑	**93.40**	87.39	93.10	92.73	91.60	92.74
ETRI-TST-HE			**93.96**	87.79	93.69	93.09	92.17	93.32
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**	57.42 / 56.59
	KoBEST BoolQ		51.34 / 51.16	50.89 / 50.54	51.88 / 50.45	50.54 / 50.54	**52.86** / **54.92**	50.54 / 50.63
Long-Form	KCSAT		**83.53** / **84.71**	34.12 / 40.00	52.90 / 47.10	69.41 / 72.94	81.18 / 78.82	55.29 / 51.76
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	69.07 / 67.13
Vicuna			79.64 / 78.43	80.21 / 80.00	**82.14** / **81.79**	67.79 / 69.50	76.43 / 73.29	76.72 / 75.08
OpenHermes			86.54 / 85.19	84.62 / 85.96	**89.42** / **89.62**	69.10 / 69.62	82.69 / 80.71	79.74 / 78.72
Alpaca			84.06 / 83.04	82.97 / 83.36	**90.58** / **90.58**	72.90 / 72.46	85.94 / 86.88	78.26 / 77.68

Table 15: Best prompt performance across ASR, ST, SQA, and SIF. For each model and task, we evaluate four prompt variants and report only the best score among them. ASR, SQA, and SIF are evaluated on clean and noisy audio when applicable. KsponSpeech-Clean/Other denote different subsets. 100 ↑ indicates CER > 100%. Blue indicates noisy results.

Dataset		Metric	Qwen3-omni	Gemma-3n	GPT-audio	Voxtral	Gemini-flash	Raon
SCA-QA								
K-history (Before)		Text-only Accuracy ↑	55.45	48.51	69.30	41.58	**79.21**	52.48
K-history (After Chosun)			79.27	71.95	79.30	51.22	**87.80**	65.85
K-sports			55.68	53.41	69.30	27.27	**78.41**	55.68
K-pop			67.96	68.93	69.90	42.72	**89.32**	55.34
K-history (Before)		SCF Score ↑	**94.64**	67.35	61.40	85.71	66.25	38.78
K-history (After Chosun)			**92.31**	45.76	32.30	85.71	59.72	23.53
K-sports			93.88	76.60	39.30	**95.83**	86.96	28.57
K-pop			**95.71**	73.24	37.50	88.64	78.26	33.33
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	40.98 / 41.28
Position - Aware	Front		91.93 / 91.93	45.34 / 44.72	72.05 / 75.78	86.34 / 85.09	**95.03** / **93.17**	43.48 / 42.03
	Front-Middle		**96.70** / **94.51**	51.65 / 53.85	83.52 / 83.52	82.42 / 85.71	89.01 / 87.91	42.70 / 44.94
	Middle-Last		**87.80** / **87.80**	47.56 / 47.56	74.39 / 70.73	76.83 / 81.71	82.93 / 85.37	32.47 / 32.47
	Last		**94.12** / 98.04	45.10 / 45.10	80.39 / 86.27	82.35 / 82.35	92.16 / **98.04**	40.00 / 38.18

Table 16: Best prompt performance on speech faithfulness tasks. For each model and task, we evaluate four prompt variants and report only the best score among them. SCA-QA reports text-only accuracy and SCF score, while PA-QA reports accuracy by evidence position. Blue indicates noisy results.

C.2 Additional Metrics for Speech Translation.

We additionally provide translation evaluation results using BLEU and METEOR. As shown in Table 14, Qwen3-Omni achieved the best performance on both ETRI-COMMON and ETRI-HE, obtaining the highest BLEU and METEOR across the compared models.

C.3 Best Prompt Reporting across Four Variants

For each task, we evaluate each model using four semantically equivalent prompt variants and report the best score among them. Therefore, Tables 15 and 16 should be interpreted as best prompt result tables, rather than prompt sensitivity or prompt average tables. This reporting protocol reduces the effect of incidental prompt wording while keeping the task objective and answer format fixed across variants.

Table 15 summarizes the best prompt performance on the general evaluation tasks, including ASR, ST, SQA, and SIF. Overall, Qwen3-omni achieves the strongest performance on Korean ASR and ST, while GPT-audio obtains the highest scores on most SIF datasets. For SQA, Gemini-flash performs best on short-form QA datasets, whereas Qwen3-omni shows stronger performance on the

EAR Example Instance

PA-QA sample `000541` *used in Figure 3.*

Context:
농구를 하는 남자 넷이 있었습니다. 그들은 야구, 미식축구, 축구는 하지 않았습니다. 그들의 이름은 세스, 태너, 헨리, 라이언이었습니다. 그들 중 한 명은 서부 지역에서 최고의 슈팅 실력을 자랑했습니다. 그는 너무 훌륭해서 거의 슛을 놓치지 않았습니다. 세상 모든 사람들은 그처럼 훌륭해지기를 원했습니다. 태너는 거의 슛을 놓치지 않는 사람이었습니다. 그는 매일 농구를 했습니다. 그는 슛을 던지고, 드리블하고, 매우 빠르게 달릴 수 있었습니다. 하지만 덩크슛은 할 수 없었습니다. 그는 농구를 너무 잘해서 훗스터스, 슈터스, 드리블러스, 덩커스 같은 팀들이 그를 영입하려고 했습니다. 그는 팀을 선택하는 데 매우 어려움을 겪었습니다. 그는 빠르게 팀을 선택해야 했습니다. 그들은 선수가 필요했기 때문에 태너는 팀을 선택해야 했습니다. 그는 훗스터스를 선택했습니다. 그들은 그의 가장 친한 친구들이었습니다. 태너는 그들과 함께 많은 경기를 치렀고, 심지어 그들의 에이스 선수였습니다. 그는 농구를 정말 즐겼고, 경기를 하는 것이 매우 즐거웠습니다. 그는 너무 즐거워서 오랫동안 농구를 했습니다.

Prompt_ko: 태너는 무엇을 할 수 없었을까요?

Choices_ko:
(A) 빠르게 달린다 (B) 덩크슛을 한다 (C) 슛을 한다 (D) 드리블한다

Ground truth: (B)

Evidence: 하지만 덩크슛은 할 수 없었습니다. (position: 47%, sentence 10/19)

Non-evidence (avg): 나머지 18개 문장의 평균

Figure 4: Representative PA-QA instance used for EAR analysis. The annotated evidence sentence directly supports the correct answer, while the remaining sentences are treated as non-evidence for comparison.

long-form KCSAT subset.

Table 16 summarizes the best prompt performance on the faithfulness-oriented evaluation tasks, including SCA-QA and PA-QA. In SCA-QA, Gemini-flash achieves the highest text-only accuracy, but Qwen3-omni shows stronger SCF scores in most domains, indicating better adherence to the provided speech context when it conflicts with prior knowledge. In PA-QA, Qwen3-omni achieves the best overall accuracy and remains competitive across evidence positions, suggesting strong long-context speech understanding under the position-aware evaluation setting.

C.4 EAR Qualitative Analysis Example

Figure 4 shows a representative PA-QA instance. The example consists of a Korean speech context, a multiple-choice question, the ground-truth answer, and an annotated evidence sentence with its relative position in the context. The evidence sentence directly supports the correct option, while the remaining sentences are treated as non-evidence, enabling us to compare model attribution to answer-relevant information against attribution to irrelevant context.

D Lists of Prompts

This appendix provides the full set of prompts used in KoALa-Bench. We include the evaluation prompts for all benchmark tasks, the LLM as judge prompt used for automatic response quality assessment, and the prompts used in the SCA-QA construction pipeline. These prompt lists are provided to make the evaluation and dataset construction procedures transparent and reproducible.

D.0.1 Prompts for LLM as Judge.

We use an LLM as judge to automatically assess response quality for tasks that require open ended generation. Figure 5 presents the prompt template used for SIF. The judge receives three fields: the input question $\{question\}$, the reference answer $\{reference\_answer\}$, and the model generated answer $\{generated\_answer\}$. It then compares the generated answer against the reference answer under three criteria: accuracy, complete-

Prompts for LLM as Judge.

SIF (Speech Instruction Following)
Objective: Evaluate whether the model's response correctly follows the given instruction.

Prompts:

```
다음은 질문과 두 개의 답변입니다. 생성된 답변을 참조 답변과 비교하여 평가해 주세요.

질문: {question}

참조 답변 (GT): {reference_answer}

생성된 답변: {generated_answer}

평가 기준:
1. 정확성: 생성된 답변이 질문에 올바르게 답하고 있는가?
2. 완전성: 생성된 답변이 참조 답변만큼 완전한 정보를 제공하는가?
3. 관련성: 생성된 답변이 질문과 관련이 있는가?

0.0부터 1.0 사이의 점수를 부여해 주세요. 점수만 숫자로 응답해 주세요.
```

Figure 5: Prompt template for LLM based judging in SIF evaluation. Given a question, a reference answer, and a generated answer, the judge outputs a scalar quality score based on accuracy, completeness, and relevance.

ness, and relevance. Accuracy measures whether the response correctly answers the question, completeness measures whether it provides sufficient information relative to the reference answer, and relevance measures whether the response remains aligned with the given question. To ensure a consistent and directly usable output, the judge is instructed to return only a scalar score between 0.0 and 1.0.

D.0.2 Evaluation Prompts

We use task specific prompt templates for all KoALa-Bench evaluation tasks. For each task, we evaluate each model with four semantically equivalent prompt variants and report the best performing result under the four prompt reporting protocol. Figures 6 and 7 list the prompt variants used for general speech tasks and QA based tasks, respectively.

D.0.3 SCA-QA Dataset Construction Prompts

We provide the full prompts used in each stage of the SCA-QA construction pipeline (Algorithm 1). Figures 8–13 present the prompts for question and answer generation (Stage 1), entity replacement (Stage 2), sentence transformation (Stage 3), TTS text normalization (Stage 4), distractor generation (Stage 5), and answer verification (Stage 6), respectively.

Prompts for KoALa-Bench Evaluation Tasks (Part I: ASR, ST, and SIF)

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.

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.

Figure 6: Prompt templates used for general speech evaluation tasks in KoALa-Bench. Part I includes ASR, speech translation, and speech instruction following tasks. Each task is evaluated with four prompt variants while maintaining the same task objective.

Prompts for KoALa-Bench Evaluation Tasks (Part II: SQA, PA-QA, and SCA-QA)

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

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)

Figure 7: Prompt templates used for QA-based evaluation tasks in KoALa-Bench. Part II includes SQA, PA-QA, and SCA-QA. For multiple-choice QA tasks, we use one-shot prompting and a fixed answer suffix to encourage responses in a consistent and parseable format.

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:
```

Figure 8: Prompt used for stage 1 of the SCA-QA dataset construction pipeline. Given a context passage, the model generates a factual question and a short answer grounded in the passage.

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 알려 주세요. (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. 귀주대첩, 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:

Figure 9: Prompt used for stage 2 of the SCA-QA dataset construction pipeline. Given the original answer entity, the model generates a new entity of the same category that does not appear in the original context.

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:
```

Figure 10: Prompt used for stage 3 of the SCA-QA dataset construction pipeline. The original answer entity is replaced with the newly generated entity to construct a counterfactual context.

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:
```

Figure 11: Prompt used for stage 4 of the SCA-QA dataset construction pipeline. Numerals and non-Korean expressions in the transformed context are normalized into spoken Korean forms for TTS synthesis.

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)
```

Figure 12: Prompt used for stage 5 of the SCA-QA dataset construction pipeline. Two additional distractor entities from the same category are generated and combined with the original and replaced answers to construct four-way multiple-choice options.

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:
```

Figure 13: Prompt used for stage 6 of the SCA-QA dataset construction pipeline. The generated sample is verified by checking whether the question can be correctly answered from the transformed context.