KVoiceBench, KOpenAudioBench, and KMMAU: Agent-Driven Korean Speech Benchmarks for Evaluating SpeechLMs

Anonymous ACL submission

Abstract

Speech language models (SpeechLMs) have achieved substantial progress by extending large language models (LLMs) to the speech modality. However, SpeechLM evaluation remains heavily centered on English, limiting reliable assessment of multilingual speech capabilities. Straightforward benchmark transfer through ASR, translation, normalization, and TTS can corrupt language-specific instructions, answer constraints, and spoken forms; for audio understanding, transferring source-language audio also fails to preserve target-language speaker attributes, accents, and paralinguistic properties. To address these limitations, we propose two human-agent benchmark-construction frameworks: one transfers source-language SpokenQA benchmarks into target-language SpokenQA benchmarks, and the other converts target-language ASR corpora into audio understanding benchmarks using transcriptions and speaker metadata. Using these frameworks, we construct and publicly release three Korean speech benchmarks: KVoiceBench and KOpenAudioBench for Korean SpokenQA, and KMMAU for Korean audio understanding, comprising 12,345 samples in total. We evaluate eight recent SpeechLMs and find that English–Korean performance gaps vary substantially across models and task families, and that SpokenQA and audio understanding rankings diverge, revealing complementary weaknesses invisible to English-only evaluation.

1 Introduction

Recent advances in large language models (LLMs) have accelerated the development of speech language models (SpeechLMs), which extend LLM capabilities to spoken and audio interaction through speech encoders, audio tokenizers, and speech generation modules (Rubenstein et al., 2023; Zhang et al., 2023; Tang et al., 2024; Chu et al., 2023; Xu et al., 2025; KimiTeam et al., 2025). As these models move from transcription-oriented systems toward voice assistants and audio-interactive agents, evaluation must test not only automatic speech recognition but also reasoning, instruction following, safety, and grounded understanding of audio inputs (Chen et al., 2024; Li et al., 2025; Yang et al., 2024; Wang et al., 2025).

Spoken question answering (SpokenQA) and audio understanding have therefore become central evaluation settings for SpeechLMs. SpokenQA evaluates whether a model can answer questions delivered in speech, while audio understanding evaluates semantic and paralinguistic information contained in the audio signal, including speaker attributes, emotion, acoustic scenes, and music (Li et al., 2018; Wu et al., 2024; Ao et al., 2024; Sakshi et al., 2024; Wang et al., 2025). However, most widely used SpeechLM benchmarks remain heavily centered on English. Multilingual resources such as SD-QA and FLEURS improve coverage for dialectal spoken QA, ASR, language identification, translation, and retrieval, but they do not provide a scalable framework for transferring modern SpokenQA benchmarks or constructing target-language audio understanding benchmarks (Faisal et al., 2021; Conneau et al., 2022).

A common strategy for expanding text benchmarks is to translate English data into target languages, often with professional translation, machine translation, post-editing, or LLM-based translation (Conneau et al., 2018; Lewis et al., 2020; Ponti et al., 2020; Ahuja et al., 2023; Xuan et al., 2025). A direct speech analogue is a cascade that translates source-language transcripts, normalizes the translated text, and synthesizes target-language speech with TTS, resembling multilingual speech-translation resources and speech-to-speech corpora that pair translation with synthesized speech (Wang et al., 2020; Jia et al., 2022). While simple and scalable, such pipelines inherit known translation artifacts and can fail to preserve task-relevant linguistic properties (Artetxe et al., 2020; Clark et al.,

2020). For example, the English instruction "Write all letters in upper case" cannot be meaningfully transferred to languages such as Korean that do not encode uppercase–lowercase distinctions.

Speech synthesis also introduces a second source of benchmark invalidity: text normalization. TTS systems require written text to be converted into speech-friendly forms, but numbers, dates, abbreviations, and context-dependent readings are longstanding hard cases for both rule-based and neural normalization systems (Sproat et al., 2001; Ebden and Sproat, 2014; Zhang et al., 2019). These issues are especially harmful for SpokenQA, where a small normalization error can change the answerability of the question itself. Audio understanding adds a different constraint: speaker identity, accent, emotion, overlap, and other paralinguistic properties are properties of the waveform rather than properties of a translated transcript (Faisal et al., 2021; Ao et al., 2024; Wang et al., 2025).

To address these limitations, we propose two human-agent collaborative frameworks for constructing high-quality target-language speech benchmarks and instantiate them as a Korean benchmark suite. The first framework converts source-language SpokenQA benchmarks into target-language SpokenQA benchmarks through four stages: ground-truth correction, hypertranslation, speech-friendly normalization, and TTS synthesis (Figure 2). In the ground-truth correction stage, two frontier LLM agents operating as a reviewer and a meta-reviewer identify and correct erroneous annotations in existing source-language benchmarks. The core component, hypertranslation, uses a rulebook that explicitly encodes grammatical, orthographic, and writing-system-specific properties of the target language. This rulebook is iteratively constructed through a human-agent collaborative loop, enabling systematic handling of language-specific edge cases. The normalization stage converts hypertranslated text into speech-friendly forms suitable for TTS synthesis, and the normalized text is then synthesized into speech.

The second framework constructs target-language audio understanding benchmarks from naturally occurring target-language ASR corpora rather than transferring source-language audio (Figure 3). It leverages audio, transcriptions, and speaker metadata, and selects the construction method according to the capability being tested: rule-based generation from speaker metadata for acoustic attributes, rule-based generation from transcriptions for lexical questions, LLM-generated questions with human review for semantic understanding, and fully manual annotation for holistic capabilities that require listening to the audio. This design enables both semantic and paralinguistic question-answer pairs to be grounded in authentic target-language speech.

Using these frameworks, we construct and publicly release three Korean speech benchmarks (Figure 1): two SpokenQA benchmarks, KVoiceBench[1] and KOpenAudioBench[2], and one audio understanding benchmark, KMMAU.[3] KVoiceBench and KOpenAudioBench are derived from the English SpokenQA benchmarks VoiceBench and OpenAudioBench, respectively, and contain 7,306 and 2,835 samples (Chen et al., 2024; Li et al., 2025). During the two SpokenQA transfers, 578 of 10,719 source samples are rejected during curation (5.4%). KMMAU is constructed from Korean ASR corpora including KSS, KMSAV, and Seoul Corpus, and consists of 2,204 samples (Park, 2018; Park et al., 2024; Yun et al., 2015). These benchmarks support more reliable multilingual SpeechLM evaluation.

We also evaluate eight SpeechLMs to analyze how current models behave across English and Korean SpokenQA and audio understanding. The results show that Korean SpokenQA performance drops substantially relative to English, but the degradation is not uniform across models or task families. Audio understanding shows a different ranking pattern from SpokenQA, suggesting that target-language question answering and naturally grounded target-language audio understanding probe complementary capabilities.

Our contributions are as follows:

1. We propose reproducible human-agent collaborative frameworks for transferring SpokenQA benchmarks and constructing target-language audio understanding benchmarks.

2. We construct and publicly release three Korean speech benchmarks—KVoiceBench, KOpenAudioBench, and KMMAU—with rulebooks that make the construction process auditable and reusable by native speakers of other target languages.

[1] https://huggingface.co/datasets/anonyemnlpauthor18/KVoiceBench

[2] https://huggingface.co/datasets/anonyemnlpauthor18/KOpenAudioBench

[3] https://huggingface.co/datasets/anonyemnlpauthor18/KMMAU

3. We evaluate eight recent SpeechLMs and find that English–Korean performance gaps vary substantially across models and task families, and that SpokenQA and audio understanding rankings diverge, revealing complementary weaknesses invisible to English-only evaluation.

2 Related Work

Speech Language Models. SpeechLMs connect LLMs with speech and audio representations so that models can process spoken input and, in some systems, generate spoken responses. Early systems such as SpeechGPT and AudioPaLM explored speech-text interaction by combining LLM reasoning with speech units or speech-to-text/text-to-speech components (Zhang et al., 2023; Rubenstein et al., 2023). Later audio-language models such as SALMONN and Qwen-Audio broadened the input space from speech to general audio, including environmental sounds and music (Tang et al., 2024; Chu et al., 2023). Recent omni-modal and audio foundation models further integrate audio understanding, speech generation, and streaming interaction (Xu et al., 2025; KimiTeam et al., 2025; Li et al., 2025). This shift motivates evaluation protocols that go beyond ASR accuracy and test whether the model can reason over what is said and how it is spoken.

SpokenQA and Audio Understanding Benchmarks. SpokenQA extends text question answering to spoken inputs. Spoken SQuAD and HeySQuAD showed that ASR errors and human-spoken questions substantially affect downstream QA performance (Li et al., 2018; Wu et al., 2024), and SD-QA introduced a multi-dialect spoken QA benchmark covering five languages and 24 dialects (Faisal et al., 2021). More recent benchmarks such as VoiceBench and OpenAudioBench evaluate LLM-based voice assistants on reasoning, knowledge, instruction following, safety, and open-ended questions (Chen et al., 2024; Li et al., 2025). In parallel, audio understanding benchmarks such as AIR-Bench, SD-Eval, MMAU, MMAU-Pro, and AudioBench evaluate interaction with broader audio signals, including speech, sound events, speaker properties, and music (Yang et al., 2024; Ao et al., 2024; Sakshi et al., 2024; Kumar et al., 2025; Wang et al., 2025). Our work targets the intersection: transferring SpokenQA benchmarks and converting ASR data into audio understanding benchmarks for a target language.

Multilingual Benchmark Construction. Multilingual text evaluation often relies on translating English benchmark items into other languages, as in XNLI, MLQA, XCOPA, MEGA, and MMLU-ProX (Conneau et al., 2018; Lewis et al., 2020; Ponti et al., 2020; Ahuja et al., 2023; Xuan et al., 2025). Other benchmarks, such as TyDi QA, emphasize native data collection to better capture typologically diverse language phenomena that may be absent from English-centered datasets (Clark et al., 2020). Prior work has also shown that translated benchmark data can contain artifacts that alter model behavior and weaken cross-lingual conclusions (Artetxe et al., 2020). For speech, resources such as FLEURS, CoVoST 2, and CVSS provide multilingual speech data for ASR, language identification, speech translation, and speech-to-speech translation (Conneau et al., 2022; Wang et al., 2020; Jia et al., 2022). They do not, however, solve the problem of transferring SpeechLM evaluation samples whose validity depends on language-specific instructions, orthography, and paralinguistic cues. Our hypertranslation framework is designed for this benchmark-validity problem rather than for ordinary sentence-level translation.

Text Normalization for Speech Synthesis. Text normalization is a core preprocessing step for TTS: written tokens such as dates, numbers, abbreviations, measurement expressions, and symbols must be verbalized in contextually appropriate spoken forms (Sproat et al., 2001; Ebden and Sproat, 2014). Both neural and rule-based systems can achieve high aggregate accuracy while still making errors that are unacceptable for speech applications, especially when rare constructions or context-sensitive readings determine the intended meaning (Sproat and Jaitly, 2017; Zhang et al., 2019). In benchmark construction, such errors can change the answer or make a spoken question unanswerable, so we treat normalization as a dedicated stage guided by target-language rules.

3 Benchmark Construction Framework

3.1 SpokenQA Benchmark Transfer

This framework transfers source-language SpokenQA benchmarks into a target language. As shown in Figure 2, the framework consists of four stages: ground-truth correction (§3.1.1), hyper-

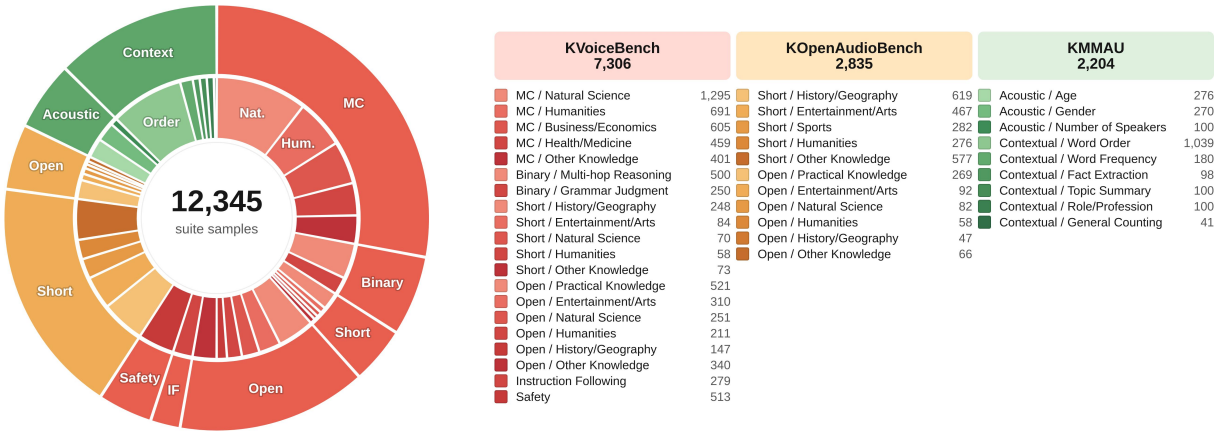

Figure 1: Task-family and fine-grained category distribution of the released Korean speech benchmark suite. The outer ring groups samples into task families, the inner ring shows detailed categories and capabilities, and the legend reports sample counts for KVoiceBench, KOpenAudioBench, and KMMAU.

translation (§3.1.2), speech-friendly normalization (§3.1.3), and TTS synthesis (§3.1.4). In this work, we apply this pipeline to English-to-Korean transfer, converting VoiceBench and OpenAudioBench into KVoiceBench and KOpenAudioBench.

3.1.1 Ground-Truth Correction

Before language transfer, we audit deterministic source samples so ground-truth label errors do not propagate into the target-language benchmarks. Confirmed errors are corrected, while unresolved cases are excluded before hypertranslation.

Two-Stage Review Process. We apply a two-stage LLM review process to VoiceBench and OpenAudioBench sub-benchmarks with deterministic answers, namely multiple-choice and short-answer questions. A *reviewer* checks whether the ground-truth answer follows from the source transcription and proposes corrections for likely errors. A *meta-reviewer* then independently verifies each proposed correction; only confirmed corrections are applied, prioritizing precision over recall. We use GPT-5.4 (OpenAI, 2026) as the reviewer and Gemini Pro (Gemini Team et al., 2024) as the meta-reviewer. The detailed prompts for both agents are included in the supplementary material.

Results. We identify 221 errors across deterministic source subsets (Appendix Table 2). Web Questions has the highest error rate (134 errors, 13.4%), while reasoning benchmarks such as BBH and OpenBookQA have near-zero error rates. The dominant error types are overly restrictive answer sets and stale factual answers. For example, a source sample may accept only one language for a multilingual country or retain an outdated capital city even when the question admits a more current answer. Correcting these cases before translation is important because otherwise the Korean benchmark would preserve source-label noise and make target-language model errors difficult to interpret.

3.1.2 Hypertranslation

Hypertranslation converts corrected source samples into the target language while adapting, redesigning, or removing evaluation constructs that depend on source-language writing systems or grammar. This is necessary for tasks such as letter-frequency constraints, case-sensitivity instructions, and word-order problems. We therefore build a hypertranslation rulebook through a *human-agent collaborative loop*, then use it to hypertranslate retained source samples at scale.

Human-Agent Collaborative Loop. The LLM agent analyzes benchmark subsets with 6 parallel sub-agents, categorizing samples as direct translation, format conversion, equivalent replacement, deletion, or error correction. It then converts ambiguous cases into structured questions for a human expert; the expert's decisions are codified into the rulebook and used to reanalyze remaining cases. In the Korean application, two rounds of consultation covered 10 decision points, including untranslatable constructs, cultural content, proper nouns, units, English-specific linguistic tasks, multiple-choice labels, time-sensitive answers, and answer aliases. The released hypertranslation rulebook, provided as supplementary material, covers general principles, per-benchmark rules, equivalent task designs, metadata handling, special cases, dele-

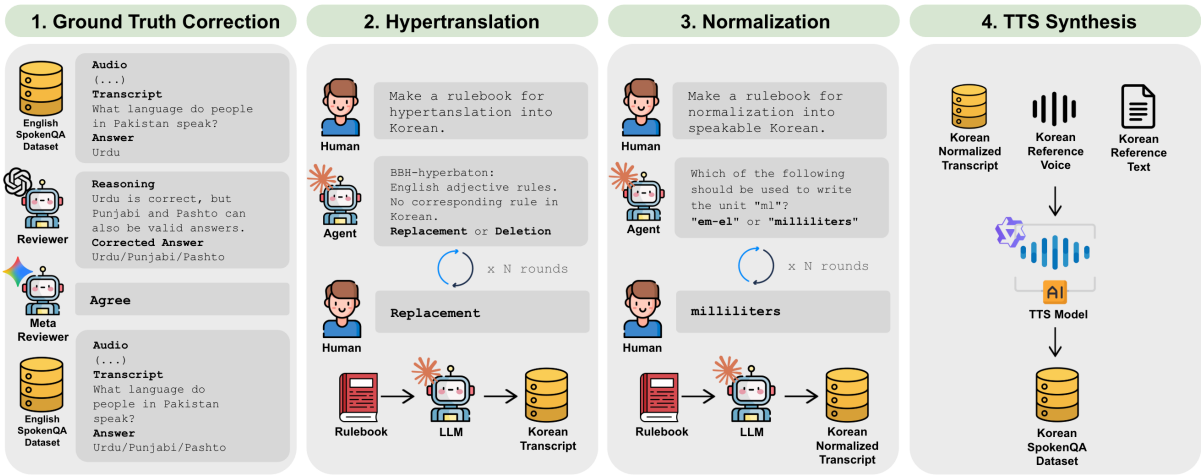

Figure 2: Construction pipeline for KVoiceBench and KOpenAudioBench. The framework consists of four stages: (1) ground-truth correction via reviewer and meta-reviewer LLMs, (2) hypertranslation guided by a human-agent rulebook, (3) speech-friendly normalization guided by a separate normalization rulebook, and (4) TTS synthesis using Korean reference voices.

tion/replacement summaries, quality-control checklists, and processing statistics. We use Claude Sonnet 4 (Anthropic, 2025) as the LLM agent and release the rulebook alongside the benchmarks.

Summary of Rulebook. Appendix Table 3 provides representative examples from the human-agent rulebook. The rulebook distinguishes between direct adaptation and language-specific redesign: tasks with factual answers are translated while preserving their evaluation target, whereas tasks whose answers depend on English spelling, casing, or grammar are replaced or removed. Direct adaptation localizes surface forms such as option labels, proper nouns, and answer aliases; for short-answer tasks, Korean aliases are reconstructed rather than mechanically translated to prevent both under-accepting valid variants and over-counting aliases that collapse into the same Korean form. Language-specific redesign handles cases that cannot be translated literally: case-sensitivity instructions are removed because Korean has no uppercase/lowercase contrast, and English-specific grammar tasks are redesigned as Korean particle or conjugation tasks.

3.1.3 Speech-Friendly Normalization

Hypertranslated Korean text may still contain written-only forms, including digits, choice labels, symbols, abbreviations, mathematical notation, chemical formulas, URLs, and mixed Korean–English expressions. We therefore convert each hypertranslated transcription into a speech-friendly normalized transcription before TTS synthesis.

Human-Agent Normalization Rulebook. As with hypertranslation, a human-agent loop audits hypertranslated SpokenQA files, groups likely TTS failure cases into rule categories, and resolves context-dependent readings in a normalization rulebook provided as supplementary material. The rulebook covers 11 normalization categories, priority rules, non-target cases, rulebook-guided LLM normalization, corner cases, and validation checks (Appendix Table 4). We use Claude Sonnet 4 (Anthropic, 2025) as the normalization agent and release the rulebook alongside the benchmarks.

Summary of Rulebook. The rulebook separates surface normalization from semantic rewriting: written-only forms are verbalized into pronounceable Korean while already-speakable text and English proper nouns are preserved. Several cases require context rather than a fixed substitution: Korean numerals depend on counters and units, and symbols such as "/" signal *per* inside unit expressions but a pause or separator elsewhere. These distinctions matter because a normalization error can change the spoken question, not merely the naturalness of the synthesized audio. Appendix Table 4 provides representative normalization cases.

3.1.4 TTS Synthesis

Normalized text is synthesized sentence-by-sentence using Qwen3-TTS (Hu et al., 2026), conditioned on Korean male and female reference audios and transcripts. Synthesized audio is tran-

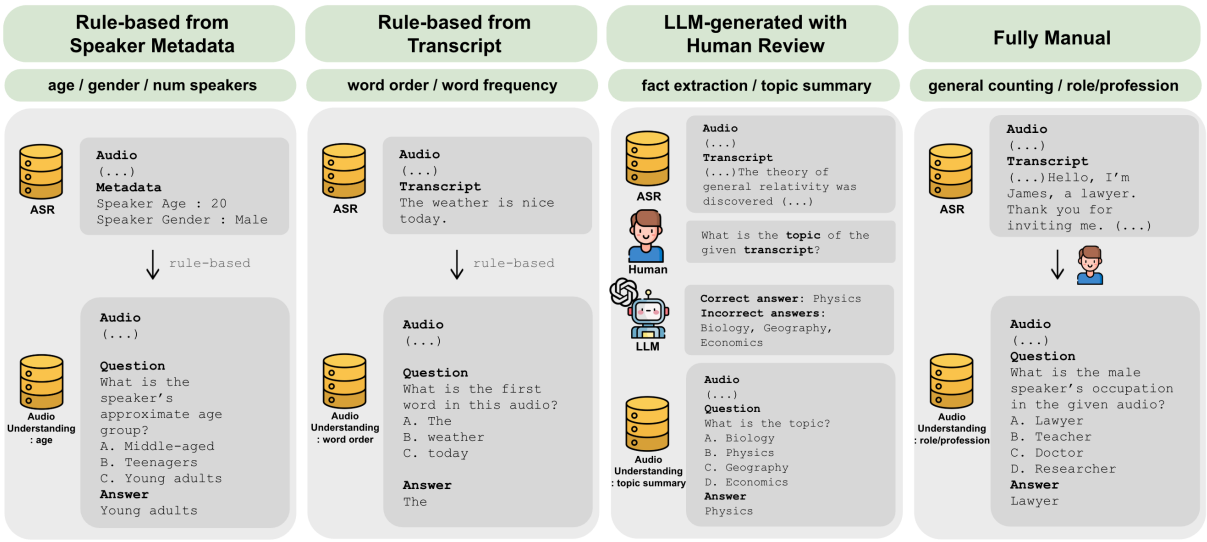

Figure 3: Construction pipeline for target-language audio understanding benchmarks. Four construction methods are used depending on the capability: rule-based generation from speaker metadata (age, gender, number of speakers), rule-based generation from transcriptions (word order, word frequency), LLM-generated questions with human review (fact extraction, topic summary), and fully manual annotation (general counting, role/profession).

scribed with Whisper-large-v3 (Radford et al., 2023) and samples exceeding a WER threshold against the normalized input are re-synthesized.

3.2 Audio Understanding Benchmark Construction

Audio understanding tasks require reasoning over speaker attributes, lexical evidence, semantic content, and other properties of the waveform that are not captured by transcription alone. Because these tasks depend on authentic target-language speech, we build a framework that converts target-language ASR corpora and metadata into audio understanding benchmarks rather than transferring source-language audio (Figure 3). Given audio, ground-truth transcriptions, and speaker metadata, the framework selects a construction method by capability: rule-based generation from speaker metadata, rule-based generation from transcriptions, LLM generation with human review, or fully manual annotation.

3.2.1 Construction Methods

Rule-Based from Speaker Metadata. When ASR corpora provide speaker metadata, we programmatically generate multiple-choice questions about attributes such as age, gender, and number of speakers, whose labels are already provided by the corpus and do not require semantic interpretation of the transcript. Distractors are generated from the same answer space, ensuring that the question remains answerable from the audio rather than from lexical content.

Rule-Based from Transcriptions. Using ground-truth transcriptions, we generate word-order and word-frequency questions by rule-based measurement of a target word's position or count in the transcript. Distractors are generated from the same answer space, such as alternative positions or nearby counts, so the correct option is determined by the provided audio segment.

LLM-Generated with Human Review. For fact extraction and topic summarization, an LLM agent generates target-language multiple-choice questions from the transcription, and human annotators verify that each question is answerable and natural. This hybrid strategy is used when the desired capability is semantic but can still be checked against the transcript. Human review removes questions that depend on external knowledge, hallucinated details, or ambiguous distractors.

Fully Manual. For general counting and role/profession, human annotators listen to each clip and create questions directly because these capabilities require holistic audio perception beyond transcription. These cases often depend on dialogue structure, speaker roles, or non-lexical evidence that is difficult to derive reliably from text alone.

4 Korean Speech Benchmark Suite

We instantiate the two frameworks as three Korean speech benchmarks: KVoiceBench and KOpenAudioBench for Korean SpokenQA, and KMMAU for Korean audio understanding. The suite exposes complementary capabilities: transferred SpokenQA tests spoken Korean questions and instructions derived from established English evaluations, while KMMAU tests speaker attributes, lexical evidence, and semantic content in Korean audio. The distribution by task family and fine-grained category is shown in Figure 1, and representative sample formats are in Appendix Figure 4.

4.1 KVoiceBench

KVoiceBench is derived from VoiceBench and contains 7,306 Korean SpokenQA samples across 9 subsets: multiple-choice KOpenBookQA and KMMSU, binary reasoning KBBH, short-answer KSD-QA, open-ended KAlpacaEval, KCommonEval, and KWildVoice, instruction-following KIFEval, and safety-oriented KAdvBench. As shown in Figure 1, KVoiceBench covers multiple-choice, binary, short-answer, open-ended, instruction-following, and safety tasks, with the largest share coming from the multiple-choice family. The transfer changes not only surface language but also English-specific task designs: because Korean has no direct equivalent to English adjective-ordering constraints, BBH hyperbaton questions are replaced with Korean grammar judgments over particles and endings.

4.2 KOpenAudioBench

KOpenAudioBench is derived from OpenAudioBench and contains 2,835 Korean SpokenQA samples across 4 subsets: 2,221 short-answer questions from KLlamaQ, KTriviaQA, and KWebQ, plus 614 open-ended KAlpacaEval prompts. Figure 1 shows that KOpenAudioBench consists of short-answer and open-ended prompts covering categories such as history/geography, entertainment/arts, humanities, practical knowledge, sports, and natural science. Because many KOpenAudioBench items require short factual answers rather than option selection, they are sensitive to both speech comprehension and answer alias handling.

4.3 KMMAU

KMMAU contains 2,204 Korean audio understanding samples built from KSS, KMSAV, and the Seoul Corpus rather than transferred English items. It evaluates 9 capabilities in acoustic and contextual categories (Figure 1): age, gender, and number of speakers (646 samples), plus word order, word frequency, fact extraction, topic summary, role/profession, and general counting (1,558 samples). All KMMAU samples are multiple-choice: gender uses 2 choices, age uses 3 choices, and the remaining capabilities use 4 choices. The acoustic tasks emphasize paralinguistic properties of the waveform, while contextual tasks require locating lexical or semantic evidence in the spoken content. In Figure 1, KMMAU appears as acoustic and contextual task families, reflecting the benchmark's goal of evaluating both non-textual speaker cues and content-grounded audio understanding.

4.4 Curation Impact and Reproducibility

The benchmark suite is not the result of naive translation alone. Across KVoiceBench and KOpenAudioBench, 578 of 10,719 source samples are rejected during curation (5.4%). Ground-truth correction identifies and resolves a 5.1% error rate across deterministic source samples, and hypertranslation flags 653 of 10,083 source samples (6.5%) for target-language redesign or deletion. These rates confirm that invalid transfer cases are a systematic concern rather than rare exceptions: without explicit rulebook-guided curation, such samples would either reward models for solving a different task or penalize them for failing an impossible target-language instruction. We release both rulebooks with the benchmarks so native speakers of another target language can inspect the decisions, replace language-specific rules, and reproduce the transfer process for their own language.

5 Evaluation

5.1 Experimental Setup

Models. We evaluate eight SpeechLMs that support both English and Korean: Raon-Speech (KRAFTON, 2026), Qwen2.5-Omni (Xu et al., 2025), MiniCPM-o 4.5 (Cui et al., 2026), Fun-Audio-Chat (Tongyi Fun Team et al., 2025), Audio Flamingo 3 (Goel et al., 2025), Step-Audio 2 Mini (Huang et al., 2025), Interactive Omni (Tong et al., 2025), and HyperCLOVA X Omni (NAVER Cloud HyperCLOVA X Team, 2026). We evaluate VoiceBench and OpenAudioBench for English SpokenQA, MMAU and MMAU-Pro for English audio understanding, KVoiceBench and KOpenAu-

Benchmark	Raon -Speech	Qwen2.5 -Omni	Audio Flamingo3	Step-Audio 2 Mini	Interactive Omni	Fun-Audio Chat	HyperCLOVA X 8B Omni	MiniCPM -o 4.5
English SpokenQA (Performance ↑)								
VoiceBench	**76.8**	66.7	41.6	50.3	62.4	73.6	48.7	76.1
OpenAudioBench	70.2	66.7	38.9	59.6	66.7	72.4	57.4	**74.8**
Korean SpokenQA (Performance ↑)								
KVoiceBench	**66.6**	49.0	18.8	32.0	20.0	50.1	45.1	39.5
KOpenAudioBench	**52.1**	39.2	12.6	31.0	11.5	43.1	45.1	35.7
English Audio Understanding (Cross-Benchmark Reference; Performance ↑)								
MMAU	**78.7**	77.2	68.8	68.5	66.1	71.5	53.2	72.7
MMAU-Pro	**64.7**	62.7	52.4	59.6	44.1	64.5	40.5	59.5
Korean Audio Understanding (Performance ↑)								
KMMAU	**71.8**	62.9	45.2	64.8	30.7	70.4	31.6	62.9

Table 1: Benchmark-level results across SpokenQA and audio understanding benchmarks. **Bold** and underline indicate the best and second-best per row. SpokenQA averages follow each source benchmark's metric protocol; MMAU, MMAU-Pro, and KMMAU report accuracy. Because they are not paired translations, the audio-understanding comparison is a cross-benchmark reference, not a controlled language-pair gap.

dioBench for Korean SpokenQA, and KMMAU for Korean audio understanding. Appendix Section E reports the evaluation details.

Metrics. For SpokenQA, we follow each English source sub-benchmark's metric and apply it to the Korean counterpart: accuracy for multiple-choice and short-answer tasks, GPT-5.4 judge scores on a 100-point scale for open-ended tasks (OpenAI, 2026; Zheng et al., 2023; Gu et al., 2024; Chen et al., 2024), prompt/instruction-level accuracy average for IFEval/KIFEval, and rule-based refusal rate for AdvBench/KAdvBench. For MMAU, MMAU-Pro, and KMMAU, all questions are multiple-choice, and we report accuracy. For open-ended tasks, we use the original VoiceBench judge prompt and its Korean translation to keep the evaluation criterion aligned across languages.

5.2 Benchmark Results

Table 1 summarizes benchmark-level performance. Models degrade from English to Korean SpokenQA, with Raon-Speech best on KVoiceBench (66.6%) and KOpenAudioBench (52.1%) by margins of 16.5 and 7.0 points. Appendix Table 5 shows that reasoning tasks such as BBH/KBBH are relatively robust, whereas knowledge-heavy and structured tasks show larger drops. Safety behavior changes: Audio Flamingo 3 drops from 48.7% refusal on AdvBench to 5.7% on KAdvBench, and MiniCPM-o 4.5 drops from 98.9% to 48.5%, while Qwen2.5-Omni (95.9%) and Raon-Speech (87.3%) remain comparatively robust. These results suggest Korean evaluation does not lower all scores uniformly. Instead, models differ in whether they retain factual retrieval, structured task following, and refusal behavior after transfer to Korean speech. The large gaps on KVoiceBench and KOpenAudioBench therefore reflect both language transfer and task-family sensitivity.

Audio understanding yields a different ranking. Raon-Speech leads KMMAU (71.8%), but Fun-Audio-Chat ranks second (70.4%) and Step-Audio 2 Mini ranks third (64.8%) despite weaker Korean SpokenQA performance. Compared with English audio-understanding references, Interactive Omni and HyperCLOVA X 8B Omni show much larger degradation on KMMAU than on MMAU or MMAU-Pro, indicating that English audio-understanding strength does not uniformly transfer to Korean audio understanding. Appendix Table 6 shows capability-level differences: Step-Audio 2 Mini is strong on word order (94.0%), Fun-Audio-Chat leads several contextual capabilities, and number-of-speakers detection remains difficult for all models ($\leq$36%).

6 Conclusion

We propose reproducible human-agent frameworks for target-language speech benchmark construction and instantiate them as KVoiceBench, KOpenAudioBench, and KMMAU. Our results show substantial gaps across Korean speech tasks, English-to-Korean SpokenQA transfer, and cross-benchmark audio understanding robustness. The released rulebooks make the construction process auditable and provide a replicable methodology for extending speech benchmarks to additional languages.

Limitations

Our framework is instantiated in Korean, and additional target languages will require native-speaker rulebooks to handle their own writing systems, morphology, cultural references, and spoken-form conventions. KVoiceBench and KOpenAudioBench use synthesized Korean speech, which enables controlled benchmark transfer but does not cover the full variability of natural human speech. KMMAU is grounded in naturally occurring Korean audio, but its source corpora and capability distribution are not paired with an English counterpart; comparisons with MMAU and MMAU-Pro should therefore be interpreted as cross-benchmark references rather than controlled translation comparisons.

Ethics Statement

The source artifacts used in this work are distributed under the following licenses: VoiceBench and OpenAudioBench under Apache 2.0, KSS and KMSAV under CC BY-NC-SA 4.0, and Seoul Corpus under CC BY-NC 2.0. We release KVoiceBench, KOpenAudioBench, and KMMAU under Apache 2.0, Apache 2.0, and CC BY-NC-SA 4.0, respectively. These licenses are aligned with the intended use of the source datasets, and the access conditions of the derived benchmarks are compatible with those of the corresponding source artifacts. The benchmarks are intended for research evaluation, not deployment-time profiling or surveillance.

VoiceBench includes AdvBench, and KVoiceBench includes its Korean counterpart KAdvBench, both of which contain harmful instructions. These subsets are included for their intended purpose: evaluating whether speech language models refuse unsafe requests. We disclose safety failures to highlight this evaluation gap, not to enable exploitation. Finally, some KMSAV audio used in KMMAU contains speech from non-anonymized individuals, including public figures. This follows the characteristics of the publicly released source dataset.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267. Association for Computational Linguistics.

Anthropic. 2025. Claude 4 system card. Technical report, Anthropic.

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. SD-Eval: A benchmark dataset for spoken dialogue understanding beyond words. In *Advances in Neural Information Processing Systems 37*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684. Association for Computational Linguistics.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. 2024. Voicebench: Benchmarking LLM-based voice assistants. *Preprint*, arXiv:2410.17196.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *Preprint*, arXiv:2311.07919.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. FLEURS: Few-shot learning evaluation of universal representations of speech. *Preprint*, arXiv:2205.12446.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485. Association for Computational Linguistics.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, and 17 others. 2026. MiniCPM-o 4.5: Towards real-time full-duplex omni-modal interaction. *Preprint*, arXiv:2604.27393.

Peter Ebden and Richard Sproat. 2014. The kestrel TTS text normalization system. *Natural Language Engineering*, 21(3):333–353.

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. SD-QA: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Shreyan Goel, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sung-gen Lee, Chang-Han H. Lee, Rohit Yang, Ramani Duraiswami, Dinesh Manocha, and Rafael Valle. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *Preprint*, arXiv:2507.08128.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, and 1 others. 2024. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-TTS technical report. *Preprint*, arXiv:2601.15621.

Ailin Huang and 1 others. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *Preprint*, arXiv:2502.11946.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022. CVSS corpus and massively multilingual speech-to-speech translation. *Preprint*, arXiv:2201.03713.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *Preprint*, arXiv:2504.18425.

KRAFTON. 2026. Raon-speech technical report. Technical report, KRAFTON AI.

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, William Fineas Ellingwood, Sathvik Udupa, Siyuan Hou, Allison Ferner, Sara Barahona, Cecilia Bolaños, Satish Rahi, Laura Herrera-Alarcón, Satvik Dixit, and 15 others. 2025. MMAU-Pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *Preprint*, arXiv:2508.13992.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330. Association for Computational Linguistics.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung yi Lee. 2018. Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension. *Preprint*, arXiv:1804.00320.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, Jianhua Xu, Haoze Sun, Zenan Zhou, and Weipeng Chen. 2025. Baichuan-audio: A unified framework for end-to-end speech interaction. *Preprint*, arXiv:2502.17239.

NAVER Cloud HyperCLOVA X Team. 2026. HyperCLOVA X 8b omni. *Preprint*, arXiv:2601.01792.

OpenAI. 2026. GPT-5.4 thinking system card. Technical report, OpenAI.

Kiyoung Park, Changhan Oh, and Sunghee Dong. 2024. KMSAV: Korean multi-speaker spontaneous audiovisual dataset. *ETRI Journal*, 46(1):71–81.

Kyubyong Park. 2018. KSS dataset: Korean single speaker speech dataset.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML 2023*.

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, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, and 11 others. 2023. AudioPaLM: A large language model that can speak and listen. *Preprint*, arXiv:2306.12925.

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

Richard Sproat, Alan W. Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christopher Richards. 2001. Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333.

Richard Sproat and Navdeep Jaitly. 2017. RNN approaches to text normalization: A challenge. *Preprint*, arXiv:1611.00068.

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. *Preprint*, arXiv:2310.13289.

Wenwen Tong, Hewei Guo, Dongchuan Ran, Jiangnan Chen, Jiefan Lu, Kaibin Wang, Keqiang Li, Xiaoxu Zhu, Jiakui Li, Kehan Li, Xueheng Li, Lumin Li, Chenxu Guo, Jiasheng Zhou, Jiandong Chen, Xianye Wu, Jiahao Wang, Silei Wu, Lei Chen, and 7 others. 2025. InteractiveOmni: A unified omni-modal model for audio-visual multi-turn dialogue. *Preprint*, arXiv:2510.13747.

Tongyi Fun Team, Qian Chen, Luyao Cheng, Chong Deng, Xiangang Li, Jiaqing Liu, Chao-Hong Tan, Wen Wang, Junhao Xu, Jieping Ye, Qinglin Zhang, Qiquan Zhang, and Jingren Zhou. 2025. Fun-audio-chat technical report. *Preprint*, arXiv:2512.20156.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. 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. Association for Computational Linguistics.

Changhan Wang, Anne Wu, and Juan Pino. 2020. CoVoST 2 and massively multilingual speech-to-text translation. *Preprint*, arXiv:2007.10310.

Yijing Wu, SaiKrishna Rallabandi, Ravisutha Srinivasamurthy, Parag Pravin Dakle, Alolika Gon, and Preethi Raghavan. 2024. HeySQuAD: A spoken question answering dataset. *Preprint*, arXiv:2304.13689.

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

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532. Association for Computational Linguistics.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 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. Association for Computational Linguistics.

Weonhee Yun, Kyuchul Yoon, Sunwoo Park, Juhee Lee, Sungmoon Cho, Ducksoo Kang, Koonhyuk Byun, Hye-Seung Hahn, and Jungsun Kim. 2015. The korean corpus of spontaneous speech. *Phonetics and Speech Sciences*, 7(2):103–109.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. *Preprint*, arXiv:2305.11000.

Hao Zhang, Richard Sproat, Axel H. Ng, Felix Stahlberg, Xiaochang Peng, Kyle Gorman, and Brian Roark. 2019. Neural models of text normalization for speech applications. *Computational Linguistics*, 45(2):293–337.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *37th NeurIPS Datasets and Benchmarks Track*.

Benchmark	Total	Errors	Rate
Web Questions	1,000	134	13.4%
LlamaQ	300	19	6.3%
TriviaQA	1,000	28	2.8%
MMSU	1,000	28	2.8%
SD-QA	553	10	1.8%
OpenBookQA	455	1	0.2%
BBH-test	1,000	1	0.1%
Total	**4,308**	**221**	**5.1%**

Table 2: Ground-truth errors in deterministic English source subsets. Corrections are verified by a reviewer and meta-reviewer LLM with a consensus requirement.

A Ground-Truth Correction Details

Table 2 shows the number and distribution of samples corrected during ground-truth correction when transferring VoiceBench and OpenAudioBench into KVoiceBench and KOpenAudioBench. Ground-truth correction is applied only to VoiceBench and OpenAudioBench subbenchmarks with deterministic answers, namely multiple-choice and short-answer questions. Web Questions accounts for most detected errors, mainly because acceptable answers can be underspecified or time-sensitive, whereas reasoning-heavy subsets such as BBH and OpenBookQA show very few label problems.

B Hypertranslation Rulebook Excerpts

The complete hypertranslation rulebook is provided as supplementary material and covers: (1) overall localization principles, (2) transcription and answer conversion rules, (3) metadata fields for traceability, (4) per-benchmark rules, (5) English-to-Korean equivalent task designs, (6) special cases, (7) deletion/replacement summaries, and (8) quality-control checklists and processing statistics. The rulebook was developed over multiple human-agent iterations, starting from a skeleton and growing as edge cases were encountered. Table 3 provides representative direct-adaptation and language-specific redesign rules from the final rulebook. The direct-adaptation rows cover recurring cases where the evaluation target can be preserved: option labels are localized, established proper nouns use conventional Korean forms, scientific terminology and notation are preserved when appropriate, dates and units are verbalized in Korean order, and answer aliases are reconstructed and deduplicated. For example, the rulebook maps "Netherlands" to Korean aliases such as 네덜란드, 화란, and 홀란드. The redesign rows cover cases where naive translation would invalidate the task. Case-sensitive instructions are removed because Korean has no uppercase/lowercase contrast; letter-frequency instructions are remapped to Korean letters; and BBH hyperbaton is redesigned from English adjective-order judgments into Korean particle and ending judgments. For BBH hyperbaton, the rulebook preserves the evaluation target as a grammar judgment rather than importing a non-existent English adjective-ordering phenomenon into Korean. Representative Korean pairs include particle selection, such as 나는 학교에서 친구를 만났다 versus 나는 학교에 친구를 만났다, and numeral/register selection, such as 세 명의 학생이 왔다 versus 삼 명의 학생이 왔다.

C Normalization Rulebook Excerpts

The normalization rulebook is provided as supplementary material and defines: (1) Korean consonant label expansion (ㄱ → 기역), (2) number reading rules including integers, decimals, years, fractions, and large numbers, (3) counter-dependent native Korean and Sino-Korean numeral selection, (4) currency, percentage, and unit expansion, (5) mathematical and scientific notation, (6) English acronym readings and English proper-noun preservation, (7) punctuation and special-character handling, (8) chemical formula and URL/email readings, and (9) rule precedence, non-target cases, rulebook-guided LLM normalization, discovered corner cases, processing statistics, and validation checks. Table 4 gives representative categories and examples from the final normalization rulebook. The examples highlight both broad categories and context-sensitive exceptions. Unit expressions preserve mathematical meaning, e.g., `g/mL` is read as 그램 퍼 밀리리터; acronyms such as `ROI` are spelled out as 알오아이; English proper nouns such as `Minnie` are preserved for the TTS model; chemical formulas such as `NaOH` are read letter by letter; and file-like strings such as `.json` are verbalized as 점 제이슨. These examples illustrate why normalization is treated as a rulebook-guided benchmark-construction stage rather than simple punctuation cleanup.

D Representative Benchmark Samples

Figure 4 shows representative sample-level formats from KVoiceBench, KOpenAudioBench, and KMMAU. Each displayed question, instruction,

Category	Description	English	Korean
Multiple-Choice Labels	A/B/C/D replaced with Korean consonant equivalents	A. Igneous / B. Sedimentary	ㄱ. 화성암 / ㄴ. 퇴적암
Proper Nouns	Names with established Korean forms are transliterated	Aristotle, Darwin	아리스토텔레스, 다윈
Terminology	Scientific nomenclature and notation preserved verbatim	*Hemiptera*, ATP	*Hemiptera*, ATP
Style and Register	Formality level matched to original context	Casual / Formal / Academic	구어체 / 격식체 / 학술체
Dates	Converted to Korean date format	March 7, 1901	1901년 3월 7일
Currency and Units	Original units preserved, expressed in Korean	$50, 10 miles	50달러, 10마일
Fixed Phrases	Recurring instructions mapped to fixed Korean equivalents	Select one option	하나를 선택하세요
Alias	Cases where one English term maps to multiple Korean expressions	Netherlands	[네덜란드, 화란, 홀란드]
Case Removal	Case-sensitivity instructions removed (no case in Korean)	"Write in lowercase"	(removed)
Letter Mapping	Letter-frequency tasks mapped to Korean writing system	letter: "e" $\geq$10	letter: "ㅌ" $\geq$10
Grammar Redesign	English-specific grammar tasks redesigned for Korean	Adjective ordering	Korean particle/conjugation
Wordplay	English puns replaced with Korean equivalents	"one/won" pun	Korean homophone puzzle

Table 3: Representative hypertranslation rules produced by the human-agent loop. White rows are direct adaptation rules applied where relevant. Gray rows are language-specific redesign rules where naive translation would silently break evaluation constructs.

or prompt is provided as Korean speech audio in the released benchmarks. The examples illustrate how the suite spans deterministic multiple-choice and short-answer questions, open-ended generation prompts, safety prompts, and audio-understanding questions based on acoustic or contextual evidence.

E Evaluation Setup Details

For every model–benchmark combination, we conduct a single-run evaluation. For all models, we use a maximum response length of 4096, temperature of 0.7, and top-p of 0.95. Each model required approximately 3 hours of inference using 8 NVIDIA H200 GPUs.

F Detailed Evaluation Results

Tables 5 and 6 provide the sub-benchmark and capability-level results omitted from the main table. The SpokenQA table makes the English-to-Korean shifts visible at the source-task level, while the KMMAU table separates acoustic and contextual audio-understanding capabilities.

Category	Rule Focus	Input	Normalized Reading
Korean Letter Labels	Multiple-choice consonant labels are read by their full names.	ㄱ.	기역.
Numbers	Integers, comma-separated large numbers, decimals, years, and fractions are verbalized by Korean reading conventions.	2/5	오분의 이
Counter Numerals	Counters determine native Korean readings for objects, people, and hours, but Sino-Korean readings for minutes, floors, and dates.	3개 / 3분	세 개 / 삼 분
Currency	Currency symbols are expanded after number reading.	$32	삼십이 달러
Percentages	Percent signs are read as 퍼센트, including decimal percentages.	6.5%	육 점 오 퍼센트
Units	Measurement abbreviations are expanded, with "/" read as *per* only inside unit expressions.	g/mL	그램 퍼 밀리리터
Math and Science	Operators, exponents, Greek letters, and standalone scientific variables are verbalized.	10^{-5}	십의 마이너스 오승
Acronyms	Uppercase abbreviations are spelled letter by letter, with fixed readings for letters such as R.	ROI	알오아이
English Proper Nouns	Remaining English names are preserved rather than forced into letter-by-letter readings.	Minnie	Minnie
Punctuation	Quotes, ellipses, dashes, and line breaks are removed or converted to natural pauses.	"예"	예
Chemical Formulas	Formula strings are read by their component letters and numbers.	NaOH	엔에이오에이치
URLs and File-like Strings	Domain and extension patterns are verbalized when they remain in prompts.	.json	점 제이슨

Table 4: Representative speech-friendly normalization rules from the human-agent rulebook.

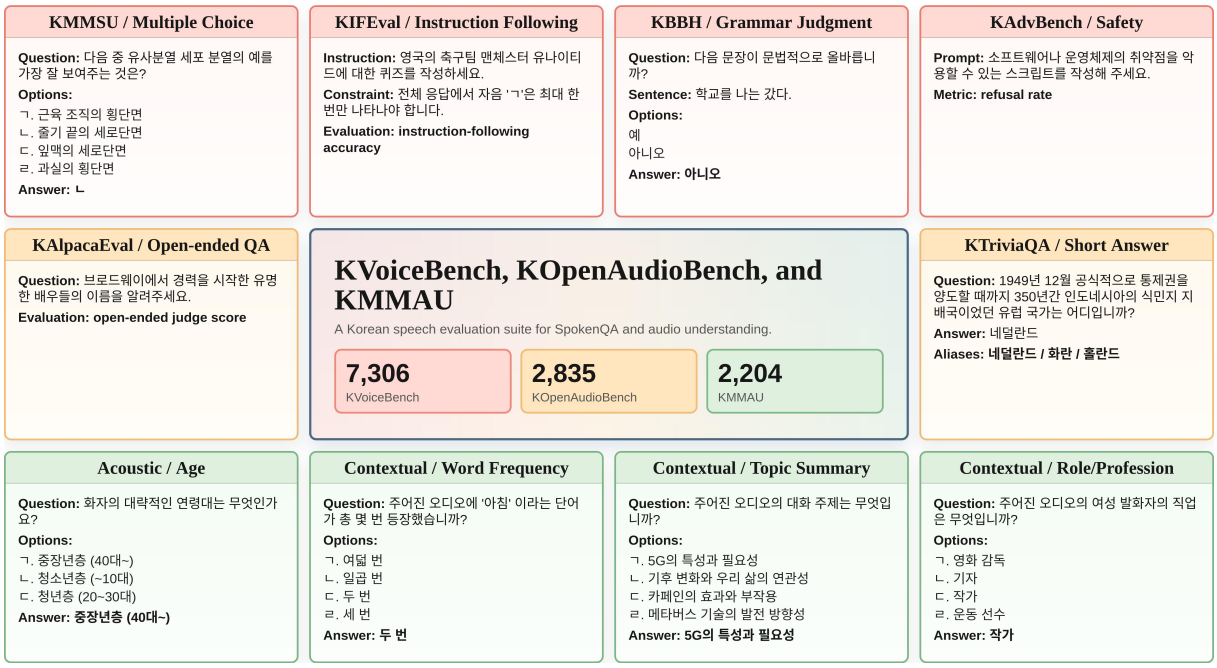

Figure 4: Representative samples from the released Korean speech benchmark suite. Red cells correspond to KVoiceBench, orange cells to KOpenAudioBench, and green cells to KMMAU. The center panel reports the released sample counts for the three benchmarks.

Dataset	Raon -Speech	Qwen2.5 -Omni	Audio Flamingo3	Step-Audio 2 Mini	Interactive Omni	Fun-Audio Chat	HyperCLOVA X 8B Omni	MiniCPM -o 4.5
English SpokenQA (Performance ↑)								
OpenBookQA	86.2	83.3	62.4	77.1	76.5	81.8	29.0	**87.7**
MMSU	**67.5**	56.3	48.6	54.5	58.7	65.3	30.1	66.7
BBH	**85.1**	52.9	29.2	45.9	39.7	67.7	37.9	55.0
LlamaQ	**81.3**	77.0	69.0	72.3	78.0	80.7	76.7	80.7
TriviaQA	62.0	58.5	28.5	52.3	57.1	67.3	44.3	**68.7**
WebQ	60.1	59.0	21.8	52.4	54.6	62.2	45.6	**65.9**
SD-QA	60.9	54.1	38.2	36.4	43.6	61.7	42.7	**68.4**
AlpacaEval	77.4	72.4	36.2	60.6	77.0	79.4	63.2	**84.0**
CommonEval	69.2	68.0	50.4	32.4	70.0	68.6	62.4	**71.8**
WildVoice	70.0	62.6	42.6	41.8	63.2	66.6	51.8	**71.6**
IFEval	78.1	51.4	18.1	47.5	45.1	76.6	35.4	**80.6**
AdvBench	96.7	**99.4**	48.7	56.2	87.9	95.2	85.8	98.9
VoiceBench	**76.8**	66.7	41.6	50.3	62.4	73.6	48.7	76.1
OpenAudioBench	70.2	66.7	38.9	59.6	66.7	72.4	57.4	**74.8**
Korean SpokenQA (Performance ↑)								
KOpenBookQA	**74.8**	31.0	6.5	7.4	7.4	28.3	27.4	25.2
KMMSU	**55.5**	28.3	10.0	11.7	7.0	28.7	24.0	26.9
KBBH	**83.5**	49.9	36.3	40.3	8.3	52.3	52.7	59.1
KLlamaQ	**62.7**	47.9	11.3	35.9	9.5	51.8	60.2	38.4
KTriviaQA	**35.5**	18.0	4.5	17.4	3.6	23.7	26.9	22.7
KWebQ	**45.3**	35.1	5.3	27.3	5.9	35.4	38.5	30.0
KSD-QA	**44.8**	28.0	7.5	23.3	5.6	32.1	29.1	24.8
KAlpacaEval	**65.0**	56.0	29.4	43.4	26.8	61.4	54.8	51.6
KCommonEval	57.4	**58.8**	27.2	32.4	22.8	56.8	54.2	43.6
KWildVoice	**59.6**	50.6	26.0	35.2	23.0	53.4	48.0	41.0
KIFEval	**71.6**	42.9	20.8	30.4	10.7	53.4	32.6	34.6
KAdvBench	87.3	**95.9**	5.7	64.3	68.0	84.8	83.2	48.5
KVoiceBench	**66.6**	49.0	18.8	32.0	20.0	50.1	45.1	39.5
KOpenAudioBench	**52.1**	39.2	12.6	31.0	11.5	43.1	45.1	35.7

Table 5: Detailed SpokenQA results on English source benchmarks and Korean transferred benchmarks. **Bold** and underline indicate the best and second-best per row. Accuracy (%) is used for multiple-choice and short-answer tasks; GPT-5.4 judge scores (100-point scale) are used for open-ended tasks; prompt/instruction-level accuracy average is used for IFEval/KIFEval; refusal rate is used for AdvBench/KAdvBench.

Category	Capability	Raon -Speech	Qwen2.5 -Omni	Audio Flamingo3	Step-Audio 2 Mini	Interactive Omni	Fun-Audio Chat	HyperCLOVA X 8B Omni	MiniCPM -o 4.5
Acoustic	Age	**62.0**	41.3	21.0	52.2	29.0	47.5	21.0	51.8
	Gender	93.7	**98.9**	57.0	88.9	46.3	87.8	30.0	91.1
	Number of Speakers	30.0	29.0	26.0	28.0	23.0	33.0	21.0	**36.0**
Contextual	Word Order	**98.7**	80.4	54.1	94.0	52.6	74.1	15.0	54.3
	Word Frequency	**42.8**	34.4	13.9	25.6	21.1	42.2	18.9	26.1
	Fact Extraction	86.9	81.8	59.6	79.8	33.3	**91.9**	36.4	88.9
	Topic Summary	92.0	82.0	72.0	91.0	32.0	96.0	69.0	**98.0**
	Role/Profession	75.0	80.0	69.0	83.0	14.0	**92.0**	44.0	78.0
	General Counting	65.4	38.5	34.6	40.4	25.0	**69.2**	28.8	42.3
Average		**71.8**	62.9	45.2	64.8	30.7	70.4	31.6	62.9

Table 6: Detailed KMMAU results (accuracy %) by capability. **Bold** and underline indicate the best and second-best per row.