CCFQA: A Benchmark for Cross-Lingual and Cross-Modal Speech and Text Factuality Evaluation

Yexing Du[1,2], **Kaiyuan Liu**[1,2], **Youcheng Pan**[2], **Zheng Chu**[1], **Bo Yang**[2], **Xiaocheng Feng**[1], **Ming Liu**[*1,2], **Yang Xiang**[*2]

[1]Harbin Institute of Technology
[2]Pengcheng Laboratory
{yxdu, mliu}@ir.hit.edu.cn, {panych, xiangy}@pcl.ac.cn

Abstract

As Large Language Models (LLMs) are increasingly popularized in the multilingual world, ensuring hallucination-free factuality becomes markedly crucial. However, existing benchmarks for evaluating the reliability of Multimodal Large Language Models (MLLMs) predominantly focus on textual or visual modalities with a primary emphasis on English, which creates a gap in evaluation when processing multilingual input, especially in speech. To bridge this gap, we propose a novel **C**ross-lingual and **C**ross-modal **F**actuality benchmark (**CCFQA**). Specifically, the CCFQA benchmark contains parallel speech-text factual questions across 8 languages, designed to systematically evaluate MLLMs' cross-lingual and cross-modal factuality capabilities. Our experimental results demonstrate that current MLLMs still face substantial challenges on the CCFQA benchmark. Furthermore, we propose a few-shot transfer learning strategy that effectively transfers the Question Answering (QA) capabilities of LLMs in English to multilingual Spoken Question Answering (SQA) tasks, achieving competitive performance with GPT-4o-mini-Audio using just 5-shot training. We release CCFQA as a foundational research resource to promote the development of MLLMs with more robust and reliable speech understanding capabilities.

Dataset — https://github.com/yxduir/ccfqa

Introduction

Large Language Models (LLMs) have achieved significant progress in recent years, driving remarkable advancements across numerous fields and applications. The popularization of Multimodal Large Language Models (MLLMs) (Li et al. 2025, 2024) has amplified the hallucination (Rawte, Sheth, and Das 2023) problem, particularly in rich multilingual scenarios. This issue is further exacerbated in cross-lingual and cross-modal settings. As shown in Figure 1, even the GPT series models struggle to mitigate hallucinations in these complex scenarios. That is, *MLLMs may yield inconsistent answers when the same factual question is asked to be answered in different languages or presented via different input modalities*.

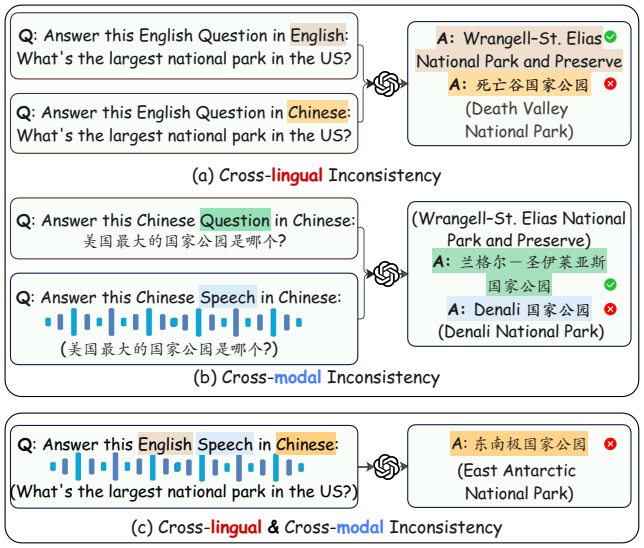

Figure 1: Factual Inconsistency in MLLMs. (a) Cross-lingual Inconsistency: inconsistent answers for the questions across different languages; (b) Cross-modal Inconsistency: inconsistent answers for the questions across different modalities; (c) Cross-lingual & Cross-modal Inconsistency.

Factuality benchmarks, such as SimpleQA (Wei et al. 2024), have gained increasing attention as effective tools for assessing hallucination in LLMs. These benchmarks use fact-based QA tasks to objectively evaluate model accuracy and reliability. However, most existing benchmarks (Zhang et al. 2025) focus on textual or visual inputs and are primarily designed for English, lacking coverage of multilingual speech scenarios. As shown in Table 1, a comprehensive benchmark for evaluating multilingual speech settings is still missing.

To bridge this gap and systematically evaluate the factual knowledge consistency of MLLMs in cross-lingual and cross-modal scenarios, we propose a new benchmark named the **C**ross-lingual and **C**ross-modal **F**actuality benchmark (**CCFQA**), which covers a total of 8 languages. The uniqueness of the CCFQA benchmark lies in the fact that each factual question is presented in both textual and spoken input forms, aiming to directly reveal whether the model's

*Corresponding author.

Benchmark (Text † / Speech ✧)	Langs.	Data Size	Modality	Metric	Open-ended	Support Task			
						QA	XQA	SQA	XSQA
TruthfulQA (Lin, Hilton, and Evans 2022)	English	817	†	Acc	✓	✓			
HaluEval (Li et al. 2023)	English	5,000	†	Acc		✓			
SimpleQA (Wei et al. 2024)	English	4,326	†	LLM	✓	✓			
Chinese SimpleQA (He et al. 2024)	Chinese	3,000	†	LLM	✓	✓			
KoLasSimpleQA (Jiang et al. 2025)	9	2,147	†	LLM	✓	✓			
SD-QA (Faisal et al. 2021)	5	11,109	✧	F1				✓	
CompA (Ghosh et al. 2024)	English	600	✧	Acc				✓	
VoiceBench (Chen et al. 2024)	English	5,783	✧	Acc / LLM	mix			✓	
SpeechIQ (Wan et al. 2025)	English	800	✧	LLM				✓	
CCFQA (ours)	8	14,400	† / ✧	F1 / LLM	✓	✓	✓	✓	✓

Table 1: Comparison of CCFQA with Existing Benchmarks. CCFQA is a cross-lingual and cross-modal factual benchmark featuring parallel speech-text question pairs across 8 languages, and supporting QA, XQA, SQA, and XSQA tasks.

internal factual knowledge exhibits bias under multilingual and multimodal inputs. CCFQA enables a systematic evaluation of MLLMs by measuring the consistency of model responses to the same question across different languages and modalities. Specifically, we collect a total of 14,400 speech and text QA samples spanning 20 distinct categories. The benchmark supports four task settings: multilingual text QA, cross-lingual text QA (XQA), multilingual spoken QA (SQA), and cross-lingual spoken QA (XSQA). Our systematic evaluation reveals that existing MLLMs exhibit notable inconsistencies in factual knowledge across languages and modalities. Even for simple questions, models often produce contradictory answers when the same query is presented in different languages or modalities, underscoring the difficulty of maintaining factual consistency with diverse inputs.

To address the challenges revealed by this benchmark and improve the factual knowledge consistency of MLLMs, we propose a novel strategy that leverages English as a pivot language to bridge the knowledge gap in cross-lingual question answering. Specifically, we design a simple yet effective end-to-end approach that transforms non-English questions into English, utilizes the strong factual reasoning capabilities of LLMs in English, and then translates the answers back into the target language. We demonstrate that this bridging strategy effectively harnesses the strengths of existing LLMs while reducing the dependency on non-English language resources, significantly enhancing the factual consistency and reliability of MLLMs.

The main contributions are summarized as follows:

- We release CCFQA, a novel benchmark for evaluating factual question answering in cross-lingual and cross-modal settings, addressing the lack of comprehensive multilingual and multimodal factuality evaluation.
- We systematically evaluate existing MLLMs and reveal inconsistencies in their answers to the factual questions across different languages and modalities, highlighting the serious challenges in maintaining factual consistency.
- We design an effective end-to-end strategy that uses English as a bridge language to leverage LLMs' strong factual knowledge, greatly improving performance and consistency in cross-lingual and cross-modal QA tasks.

Related Work

Factuality Benchmarks

Recently, a series of evaluation benchmarks (Liu et al. 2025) have emerged in the LLM field, with factuality benchmarks being one of them. Fact-based question answering is a vital research area for evaluating LLMs. TruthfulQA (Lin, Hilton, and Evans 2022), known as the first fact-based benchmark in the era of LLMs, assesses truthfulness by measuring how LLMs handle questions involving common human misconceptions. Its goal is to prevent models from generating imitative falsehoods learned from their training data. HaluEval (Li et al. 2023) addresses "hallucinations," which are coherent but factually incorrect texts, by using human-annotated samples. SimpleQA (Wei et al. 2024) employs adversarially collected, short, fact-seeking questions with single answers to test a model's factual recall. Extending fact-based evaluation to other languages, Chinese SimpleQA (He et al. 2024) is the first comprehensive Chinese benchmark with high-quality answers, mainly focusing on evaluating Chinese-centered LLMs. Finally, KoLasSimpleQA (Jiang et al. 2025) is the first multilingual fact-based benchmark, assessing nine languages on both general and language-specific knowledge.

Spoken Question Answering

Evaluating Spoken Question Answering (SQA) is an evolving field. SD-QA (Faisal et al. 2021) is a typical work addressing dialectal variations with a dataset of over 68,000 spoken prompts, enabling real-world performance and fairness evaluations. CompA (Ghosh et al. 2024), while not a direct benchmark, offers insights into multisensory fusion relevant for designing multi-modal SQA systems. VoiceBench (Chen et al. 2024) provides a comprehensive benchmark for LLM-based voice assistants, assessing general knowledge, instruction-following, and safety under realistic audio conditions. SpeechIQ (Wan et al. 2025) introduces a cognition-inspired pipeline, SIQ, which evaluates models across three levels of Bloom's Taxonomy, providing a holistic assessment of a model's understanding. All of these benchmarks focus on a single voice modality and use non-open-ended answers, highlighting a common limitation in current SQA evaluation methods.

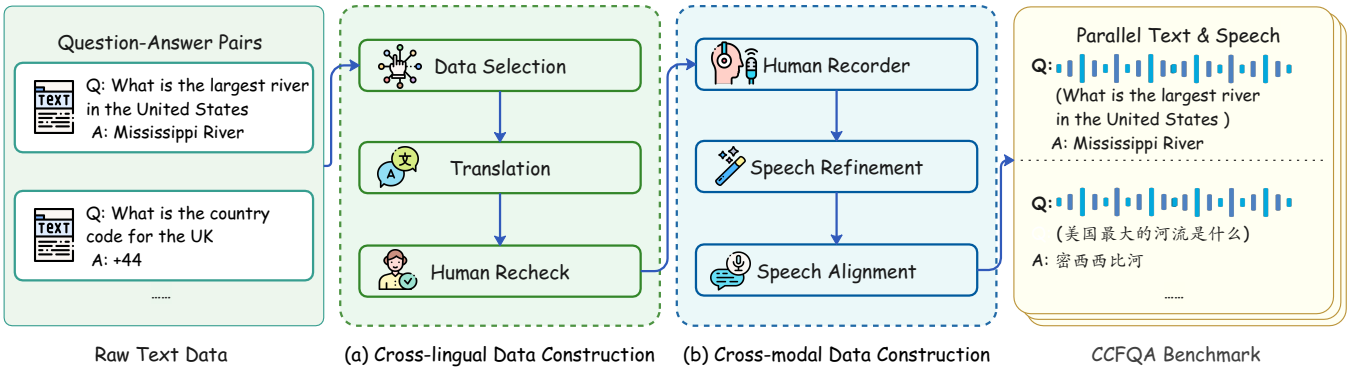

Figure 2: CCFQA Dataset Construction: (a) Cross-Lingual Data Construction, (b) Cross-Modal Data Construction.

The CCFQA Benchmark

Overview

This section briefly outlines the construction of the CCFQA benchmark, designed to evaluate MLLMs on multilingual spoken factual question answering. The process consists of two phases: cross-lingual and cross-modal data construction.

Cross-lingual Data Construction

To construct the benchmark, we leverage text-based QA datasets: MKQA (Longpre, Lu, and Daiber 2021) and MOOCCubeX (Yu et al. 2021). MKQA provides a wealth of open-domain questions, with the coverage of general knowledge topics such as movies, music, and sports. MOOCCubeX offers questions rooted in educational contexts, covering a wide array of academic subjects.

Data Selection. A critical step in our methodology was the manual curation of the initial data pool. We established a rigorous set of exclusion criteria to ensure the utmost quality. Questions were filtered and excluded if they exhibited any of the following characteristics:

- **Ambiguity:** Questions that are poorly phrased, vague, or whose answers may change over time (e.g., "Who is the current prime minister?").
- **Sensitivity:** Content containing personally identifiable information (PII), offensive language, or highly controversial political or social topics.
- **Factual Incorrectness:** Pairs where the provided answer was outdated, disputed, or verifiably incorrect.
- **Culture Independent:** Questions whose answers are not universally factual but vary across different cultural or legal contexts. (e.g., "What is the age of consent?")

Translation. Subsequently, the filtered questions and their corresponding canonical answers were professionally translated into other seven target languages. To achieve high-quality translation, we utilized the advanced machine translation capabilities of GPT-4.1. Each translation prompt was carefully engineered to preserve the factual nature of the query and the accuracy of the answer, explicitly instructing the model to maintain semantic equivalence and formal tone.

Human Recheck We performed manual verification for certain languages (Chinese, English, Japanese) and employed back-translation followed by review for the remaining languages to ensure the accuracy and consistency of all questions and answers.

Cross-modal Data Construction

The final phase involves converting the verified text-based QA pairs into a high-quality speech dataset for evaluating MLLMs' speech processing. This includes guided human recordings, audio enhancement, and ASR-based quality checks to identify and re-record low-quality samples. This iterative process ensures the speech data accurately matches the text.

Human Recording. We recruit a balanced and diverse group of native speakers (including male and female) for the 8 target languages to record the **parallel speech-text samples**. Each volunteer is instructed to read the sentences with clear, natural enunciation, at a consistent pace, and in a neutral tone, avoiding emotional inflections or disfluencies. This protocol is designed to produce high-quality, clean audio recordings ideal for MLLM evaluation.

Speech Refinement. To enforce a final layer of quality control on the audio data, we implement a systematic speech refinement loop. We apply audio augmentation to enhance samples with low volume levels. Additionally, we employ Whisper-large-v3 (Radford et al. 2023), an Automatic Speech Recognition (ASR) model, for a comprehensive quality check. The ASR system transcribes every recorded audio file. We then calculate the Word Error Rate (WER) and Character Error Rate (CER) by comparing the ASR-generated transcripts with the original ground-truth text.

Speech Alignment. Any audio file exhibiting a WER above a predefined low threshold is automatically flagged as having a potential mispronunciation or recording artifact. Our volunteers are then asked to re-record these specific flagged utterances. This iterative process of recording, ASR-based verification, and re-recording ensures that the final audio dataset is of exceptional quality, accurately reflects the underlying text, and provides a reliable basis for cross-modal evaluation for MLLMs.

Language	Code	Question (Text † / Speech ✧)	Answer †	Time (s)
- English	eng	Which philosopher proposed the idea that "the unexamined life is not worth living"?	Socrates	7.4
- Chinese	cmn	哪位哲学家提出了“未经审视的生活是不值得过的”这一观点?	苏格拉底	7.0
- French	fra	Quel philosophe a proposé l'idée que « une vie sans examen ne vaut pas la peine d'être vécue » ?	Socrate	8.3
- Japanese	jpn	「吟味されていない人生は生きるに値しない」と提唱した哲学者は誰ですか?	ソクラテス	5.8
- Korean	kor	" 성찰하지않은삶은살가치가없다" 는견해를제시한철학자는누구입니까?	소크라테스	4.9
- Russian	rus	Какой философ выдвинул идею, что «непроверенная жизнь не стоит того, чтобы её жить»?	Сократ	7.2
- Spanish	spa	¿Qué filósofo propuso la idea de que "una vida no examinada no merece ser vivida"?	Sócrates	7.4
- Cantonese	yue	哪位哲學家提出了「未經審視的生活是不值得過的」這一觀點?	蘇格拉底	6.8

Table 2: Questions and Answers in All Supported Languages for One Instance in CCFQA.

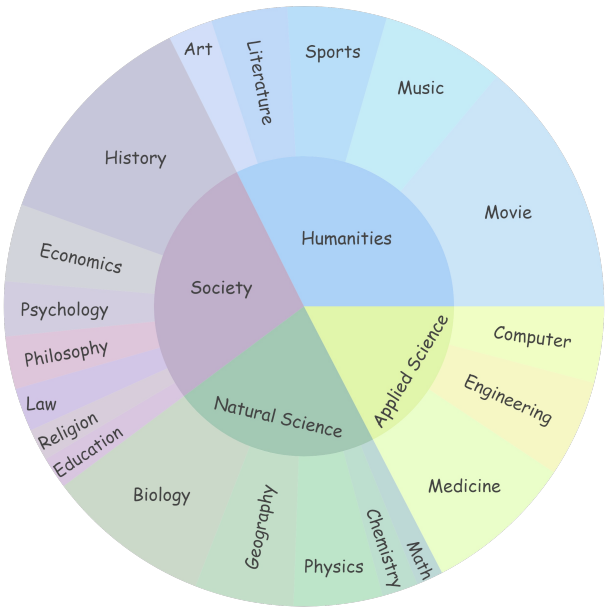

Figure 3: The Question Categories in the CCFQA.

Question Category		
Total		1800
- Applied Science	Computer	74
	Engineering	94
	Medicine	145
- Humanities	Art	43
	Literature	74
	Movie	250
	Music	120
	Sports	96
- Natural Science	Biology	160
	Chemistry	35
	Geography	96
	Math	26
	Physics	86
- Society	Economics	75
	Education	31
	History	218
	Law	42
	Philosophy	51
	Psychology	53
	Religion	31

Benchmark Statistics	
Languages	8
English Chinese French Japanese Korean Russian Spanish Cantonese (hk)	
Data Size	1800 × 8
- Validation	800 × 8
- Test	1000 × 8
Test Sample Size per Task	
- QA	1000 × 8
- XQA	1000 × 8 × 8
- SQA	1000 × 8
- XSQA	1000 × 8 × 8
Question Length	
- *maximum length*	371
- *minimum length*	6
- *avg length*	49.8
Ref. Answer Length	
- *maximum length*	59
- *minimum length*	1
- *avg length*	9.3

Table 3: Benchmark Statistics. CCFQA supports the evaluation of QA, XQA, SQA, and XSQA for MLLMs. It features parallel speech-text question pairs across 8 languages. Due to the cross-lingual evaluation settings in XQA and XSQA, a full evaluation of one MLLM requires 128,000 individual requests.

Benchmark Statistics

This section introduces the CCFQA benchmark, a novel and comprehensive resource designed for evaluating the knowledge consistency of MLLMs across various QA tasks. Specifically, CCFQA supports the evaluation of QA, XQA, SQA, and XSQA for MLLMs, making it a versatile tool for multimodal and multilingual research.

Language & Domain Distribution. As illustrated in Figure 3, our dataset is meticulously constructed from content spanning a total of 8 languages with 4 major domains (humanities, society, natural science and applied science), which are further categorized into 20 sub-domains.

Test & Validation Set Composition. The CCFQA benchmark contains a total of **1800** parallel speech-text question pairs covering 8 languages (each question available in both text and speech forms). Among them, **800** pairs are used for validation and **1000** pairs for testing. Due to the cross-lingual evaluation settings in XQA and XSQA, a full evaluation of one MLLM requires **128,000** individual requests.

Cross-Lingual and Cross-Modal Consistency

The CCFQA benchmark is designed to evaluate the consistency of MLLMs across different languages and input modalities. A robust MLLM should give the same answer to an identical factual question, regardless of the input language or whether the question is presented as text or speech. However, most existing benchmarks lack the fully parallel data needed for such an evaluation, especially in multilingual spoken scenarios. To address this gap, CCFQA provides a fully parallel speech-text dataset covering eight languages. This enables systematic assessment of:

- **Cross-lingual Consistency:** Can the model produce equivalent answers across multiple languages?
- **Cross-modal Consistency:** Can the model maintain answer quality across text and speech inputs?

Thanks to this parallel design, evaluations can be controlled and directly compared, revealing modality and language biases that are often hidden in non-parallel benchmarks.

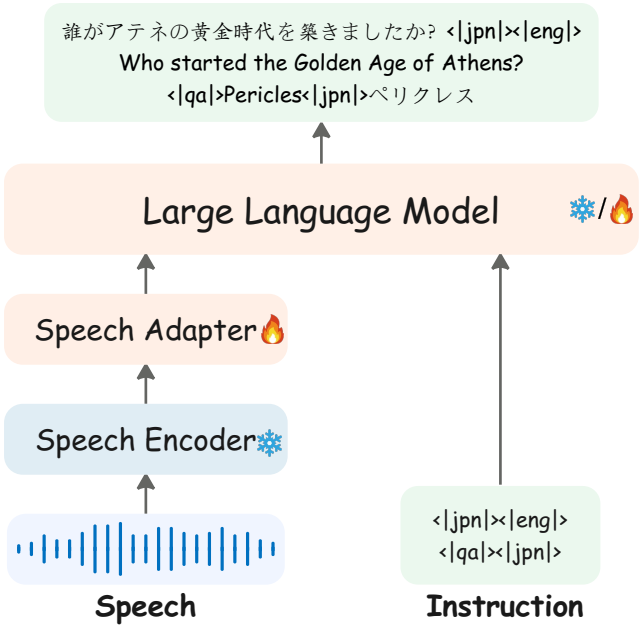

Figure 4: The Architecture of LLM-SQA.

Few-Shot Transfer Learning for LLM-SQA

Pretraining. Following the design of LLM-SRT (Du et al. 2025), we adopt a sequential curriculum learning strategy that trains the model on three tasks in order: (1) Automatic Speech Recognition (ASR), (2) Speech Recognition and Translation (SRT), and (3) Spoken Question Answering (SQA). To help the model distinguish between different tasks, we design explicit task instructions that ensure clear separation between instructions and predicted outputs, as shown in Figure 4. The ASR and SRT tasks are trained on the FLEURS dataset (Conneau et al. 2023).

SQA Few-shot Learning. We adopt a two-stage training strategy for the SQA task. In the first stage, we perform supervised fine-tuning using approximately 3,000 synthetic English speech samples paired with corresponding text. This enables the model to learn the task structure and factual question-answering ability in a high-resource setting.

In the second stage, we enable cross-lingual transfer by applying a **5-shot** few-shot learning setup on target languages. With only a handful of annotated examples, the model quickly adapts to new languages and modalities. By leveraging English as a bridge, the learned knowledge is effectively transferred to other languages (e.g., Japanese or Cantonese), greatly reducing the reliance on large-scale annotated data in each target language. This strategy enables our model to perform cross-lingual spoken question answering across **eight languages**, despite minimal supervision.

Language and Task Support. Our MLLM is pre-trained from scratch based on GemmaX2-9B (Cui et al. 2025), with added support for Cantonese (Hong Kong SAR). Our model supports ASR, Speech-to-Text Translation (S2TT), and SQA tasks together within a single query.

Task	Instruction	Prediction
ASR	<|eng|>	what is the largest city in the uk
	<|fra|>	quelle est la plus grande ville du royaume-uni
	<|spa|>	cual es la ciudad mas grande de uk
SRT	<|eng|><|fra|>	what is the largest city in the uk<|eng|><|fra|> quelle est la plus grande ville du royaume-uni
	<|spa|><|eng|>	cual es la ciudad mas grande de uk<|spa|><|eng|> what is the largest city in the uk
SQA	<|qa|>	what is the largest city in the uk <|qa|>London
	<|spa|><|eng|> <|qa|>	cual es la ciudad mas grande de uk<|spa|><|eng|> what is the largest city in the uk <|qa|>London
	<|spa|><|eng|> <|qa|><|fra|>	cual es la ciudad mas grande de uk<|spa|><|eng|> what is the largest city in the uk <|qa|>London<|fra|>Londres

Table 4: Instruction Design. LLM-SQA consists of a speech encoder, speech adapter, and LLM. A curriculum learning strategy sequentially trains the ASR, SRT, and SQA tasks. We train the <|qa|> task in English, and for other languages, cross-lingual knowledge transfer can be achieved with only 5-shot trainings, using English as a bridge.

Experiments

Experiment Setting

Model Architecture. Our MLLM includes a frozen speech encoder, a trainable adapter composed of a Q-Former with 80 queries of dimension 768 and an MLP, and an LLM, as detailed in Table 5.

Training Details. Experiments run for one week on four A100 (80GB) GPUs. We use AdamW with a peak learning rate of 1×10^{-4}, 1000-step warm-up, and linear decay thereafter.

Comparing MLLMs. We compare five MLLMs: GPT-4o-mini (OpenAI 2023), Phi-4-Multimodal (Abouelenin et al. 2025), Qwen2-Audio (Chu et al. 2024), and the Qwen2.5-Omni series (Xu et al. 2025).

Evaluation Metrics. We evaluate using the F1 score and an LLM judge. Details on the judge selection and experimental setup are in the Appendix.

Modules	Param	Training stage	Details
Speech Encoder	~635M	-	Whisper's encoder
Speech Adapter	~80.5M	all	Q-Former and MLP
LLM	~9.2B	-	GemmaX2-9B
LLM adapter	~8.9M	II&III	LoRA (r=16, alpha=32)
Total	~10B		

Table 5: MLLM Training Settings. The blue color indicates the number of trainable parameters.

F1 / LLM Acc ↑	**CCFQA** (Text † / Speech ✧) Cmn	Eng	Fra	Jpn	Kor	Rus	Spa	Yue	Avg.
				† QA (X→X)					
GPT-4o-mini	**59.2 / 63.6**	**78.7 / 82.0**	**74.1 / 73.7**	**60.3 / 63.4**	**50.9 / 55.3**	**62.7 / 42.0**	**74.1 / 76.3**	**51.6 / 59.0**	**63.9 / 64.4**
Phi-4-Multimodal	6.9 / 13.4	32.4 / 28.7	32.6 / 15.8	6.5 / 11.9	7.1 / 5.6	22.4 / 11.3	32.0 / 16.6	4.0 / 6.7	18.0 / 13.8
Qwen2-Audio	17.7 / 40.5	46.5 / 54.4	46.0 / 37.2	9.9 / 18.6	10.3 / 11.4	37.3 / 18.2	45.8 / 39.9	9.7 / 24.4	27.9 / 30.6
Qwen2.5-Omni-3B	13.4 / 18.1	29.7 / 26.7	33.0 / 13.9	7.0 / 2.0	4.7 / 2.5	27.1 / 4.3	37.0 / 12.8	13.2 / 13.8	20.6 / 11.8
Qwen2.5-Omni-7B	45.3 / 53.2	66.4 / 60.3	58.1 / 39.7	38.1 / 32.9	25.1 / 19.9	48.1 / 22.7	58.9 / 43.7	31.7 / 33.2	46.5 / 38.2
				† XQA (X→8)					
GPT-4o-mini	**60.3 / 63.1**	**64.4 / 68.8**	**62.1 / 65.0**	**57.2 / 59.4**	**56.1 / 56.6**	**59.2 / 58.7**	**61.7 / 65.8**	**56.6 / 60.0**	**59.7 / 62.2**
Phi-4-Multimodal	17.8 / 15.5	19.1 / 19.9	18.5 / 16.0	17.8 / 15.2	17.7 / 10.0	17.6 / 15.7	18.6 / 16.4	16.8 / 13.3	18.0 / 15.2
Qwen2-Audio	25.5 / 22.3	25.1 / 21.9	24.1 / 18.6	23.2 / 18.1	23.1 / 13.9	24.2 / 17.4	24.3 / 19.9	23.9 / 21.1	24.2 / 19.1
Qwen2.5-Omni-3B	12.5 / 11.1	18.6 / 16.9	17.9 / 10.7	8.4 / 2.8	10.5 / 3.6	13.6 / 6.2	18.4 / 10.6	11.6 / 9.4	13.9 / 8.9
Qwen2.5-Omni-7B	45.7 / 42.4	47.9 / 39.7	43.0 / 35.8	41.7 / 32.7	37.4 / 26.9	40.7 / 31.7	43.5 / 35.4	36.5 / 31.4	42.0 / 34.5
				✧ SQA (X→X)					
GPT-4o-mini-Audio	38.6 / 36.6	**75.9 / 74.7**	54.0 / **40.4**	42.9 / **39.5**	31.8 / 28.9	53.9 / 28.2	63.7 / **57.1**	21.1 / 17.8	47.7 / **40.4**
Phi-4-Multimodal	7.1 / 25.3	40.0 / 56.5	39.1 / 34.4	9.1 / 18.9	1.7 / 1.2	10.5 / 0.4	38.3 / 37.7	2.4 / 1.6	18.5 / 22.0
Qwen2-Audio	19.4 / 31.5	48.6 / 31.3	44.7 / 19.2	8.1 / 5.0	8.9 / 2.6	38.0 / 5.6	46.1 / 24.6	8.0 / 16.3	27.7 / 17.0
Qwen2.5-Omni-3B	41.3 / 39.6	59.1 / 41.4	35.7 / 15.3	21.1 / 11.4	17.2 / 7.9	36.9 / 9.3	48.1 / 23.2	19.4 / 19.1	34.9 / 20.9
Qwen2.5-Omni-7B	**55.2 / 53.5**	68.2 / 58.5	43.2 / 25.7	33.0 / 25.7	21.4 / 11.5	46.0 / 21.6	54.6 / 37.2	**30.7 / 31.7**	44.0 / 33.2
LLM-SQA (ours)	51.3 / 45.5	74.8 / 60.5	**60.3** / 36.8	**43.4** / 38.8	**41.6 / 36.1**	**54.1 / 32.5**	**64.5** / 45.3	25.9 / 27.0	**52.0** / 40.3
				✧ XSQA (X→8)					
GPT-4o-mini-Audio	44.9 / 34.8	**62.9 / 63.7**	42.7 / 36.8	46.1 / 37.3	40.7 / 29.9	48.0 / **39.9**	53.0 / **50.0**	26.9 / 16.5	45.7 / 38.6
Phi-4-Multimodal	20.7 / 6.7	23.3 / 10.4	22.6 / 10.7	20.6 / 5.6	21.5 / 0.6	18.4 / 0.2	22.3 / 10.6	18.7 / 0.5	21.0 / 5.7
Qwen2-Audio	25.9 / 16.6	26.5 / 17.2	23.5 / 10.2	23.7 / 8.1	21.2 / 2.3	23.2 / 5.8	24.4 / 12.3	24.4 / 13.6	24.1 / 10.7
Qwen2.5-Omni-3B	34.6 / 24.1	42.6 / 27.3	26.1 / 13.3	26.6 / 13.3	25.8 / 12.9	26.7 / 12.9	35.1 / 19.3	20.6 / 13.9	29.8 / 17.1
Qwen2.5-Omni-7B	47.6 / 38.8	48.1 / 45.3	32.8 / 23.5	39.3 / 26.2	34.5 / 21.0	37.7 / 24.9	42.1 / 33.6	26.1 / 22.6	38.5 / 29.5
LLM-SQA (ours)	**56.6 / 45.8**	48.9 / 40.3	**49.9 / 38.6**	**52.2 / 39.9**	**52.2 / 40.9**	**52.3** / 39.6	**54.4** / 44.9	**44.9 / 27.7**	**51.4 / 39.7**

Table 6: F1 and LLM-based accuracy of MLLMs on the CCFQA benchmark, including cross-lingual (QA → XQA / SQA → XSQA) and cross-modal (QA → SQA / XQA → XSQA) settings. Performance degradation is observed in both cross-lingual and cross-modal scenarios.

Main Results

As shown in Table 6, we evaluate MLLMs across linguistic (monolingual vs. cross-lingual) and modalities (text vs. speech) settings, with a focus on consistency in factual knowledge. We have the following insightful observations:

Text-based QA and XQA. In text-based tasks, GPT-4o-mini achieves the best performance. Among open-source MLLMs, Qwen2.5-Omni-7B obtains the highest score. In terms of language-wise performance, English (eng), French (fra), and Spanish (spa) achieve relatively high scores, while Korean (kor) and Cantonese (yue) perform relatively worse.

Speech-based SQA and XSQA. In speech-based tasks, GPT-4o-mini-Audio achieves the best performance in English (eng), while Qwen2.5-Omni-7B performs best in Mandarin (cmn) and Cantonese (yue). LLM-SQA demonstrates strong overall performance through a knowledge transfer strategy based on few-shot learning.

Comparison Between F1 / LLM Scores. We observe differences between F1 and LLM-based accuracy scores. Low F1 with high accuracy suggests the MLLM knows the facts but struggles to follow instructions (prompt requires answers without explanations). Conversely, high F1 but low accuracy indicates fluent, content-rich answers that are related to the question but factually incorrect, reflecting hallucinations.

Consistency ↑	**Cross-Lingual**	**Cross-Modal**
GPT-4o-mini	**96.6 / 95.5**	62.7 / 62.1
Phi-4-Multimodal	90.8 / 25.9	62.7 / 37.5
Qwen2-Audio	62.4 / 62.9	55.6 / 56.0
Qwen2.5-Omni-3B	75.4 / 81.8	56.5 / 52.0
Qwen2.5-Omni-7B	90.3 / 87.2	**90.3 / 85.5**

Table 7: Cross-Lingual and Cross-Modal Consistency. Scores are the ratio of the performance across the four tasks.

Cross-Lingual Challenge. As shown in Table 7, in cross-lingual settings (XQA and XSQA), most models (except for the GPT-4o-mini) exhibit a significant performance drop compared to their results on QA and SQA tasks. This highlights the difficulty of multilingual alignment, as models generally achieve their best performance in English but perform poorly on low-resource languages.

Cross-Modal Challenge. As shown in Table 7, a pronounced performance degradation is observed across most models when the input modality shifts from text to speech. This "modality gap" highlights the inherent challenges in multimodal alignment. Notably, Qwen2.5-Omni-7B exhibits superior cross-modal consistency, highlighting the effectiveness of the Omni design.

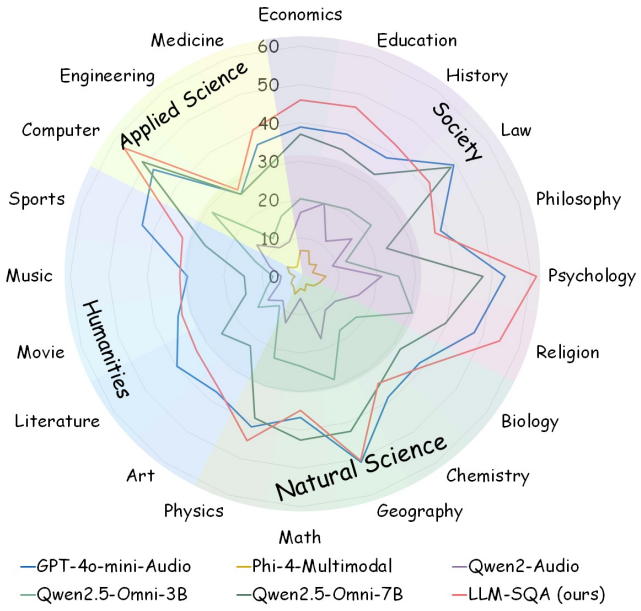

Figure 5: Performance Across Categories on XSQA Task.

Further Analysis

Category Performance on XSQA. Figure 5 presents a radar chart illustrating model performance across 20 knowledge categories on the XSQA task. We observe that LLM-SQA demonstrates the most balanced and consistently high performance across both scientific (e.g., Biology, Chemistry, Medicine) and social domains (e.g., Psychology, History, Education), significantly outperforming baselines. In contrast, other models such as Qwen2.5-Omni-7B show more varied performance, excelling in specific areas like Computer and Law but underperforming in fields such as Movie, Music and Philosophy.

Speech Length Analysis. We evaluate SQA performance across different speech durations to assess robustness to input length. LLM-SQA achieves the highest average accuracy (40.3), surpassing all open-source baselines and matching GPT-4o-mini-Audio (40.4). It performs especially well on short (0–5s) and medium (5–10s) segments, where models like Qwen2-Audio and Phi-4-Multimodal show notable drops. This demonstrates LLM-SQA's strength in handling brief spoken queries common in real-world use. In contrast, while Qwen2.5-Omni-7B performs decently overall (34.5), it lags behind LLM-SQA across all length ranges, highlighting the effectiveness of our instruction-tuning strategy.

Language Tags for MLLMs. Most current MLLMs are built upon base LLMs and extend their capabilities to support multimodal inputs by incorporating special tokens. The design of these special tokens plays a crucial role in effectively distinguishing different languages and tasks. For example, phi-4-multimodal employs a special token sequence for SQA such as `<|user|><|audio_1|><|end|><|assistant|>`. While this design works well for standard SQA, it struggles with XSQA, causing a sharp drop in performance. This shows that the design of special tokens needs to carefully consider language support.

LLM Acc ↑	0-5s	5-10s	10-30s	Avg.
✧ SQA (X→X)				
GPT-4o-mini-Audio	38.6	**42.6**	**40.1**	**40.4**
Phi-4-Multimodal	17.7	26.4	27.4	22.0
Qwen2-Audio	14.7	19.5	18.6	17.0
Qwen2.5-Omni-3B	20.5	21.7	18.6	20.9
Qwen2.5-Omni-7B	32.1	34.9	29.0	33.2
LLM-SQA (ours)	**40.3**	40.8	36.1	40.3

Table 8: LLM-based Accuracy Across Speech Lengths.

WER ↓				CER ↓			
Eng	Fra	Rus	Spa	Cmn	Jpn	Kor	Yue
3.2	13.8	18.2	9.9	6.8	8.7	8.2	16.8

Table 9: Speech Quality Check. We use Whisper-Large-v3 to evaluate ASR error rates.

Speech Quality Check. The table 9 shows the Word Error Rate (WER) and Character Error Rate (CER) in our benchmark. English (Eng) and Mandarin (Cmn) perform best, with WER and CER of 3.2 and 6.8. Some languages like French (Fra), Russian (Rus), and Cantonese (Yue) have higher error rates mainly due to many domain-specific proper nouns that are harder for ASR model to recognize, even when pronounced correctly. Overall, these results indicate that the speech data in our benchmark is of high quality with generally low error rates across languages.

Conclusion

This paper proposes a factual benchmark, CCFQA, for evaluating the consistency of MLLMs in cross-lingual and cross-modal scenarios. The benchmark covers 8 languages and consists of parallel, real human speech and text data. In addition, we introduce a simple yet effective few-shot learning strategy that leverages English as a bridge language for cross-lingual knowledge transfer.

Limitations

Although this paper explores cross-lingual and cross-modal consistency between speech and text inputs, the current benchmark remains limited to these two modalities. Future work could consider extending it to additional modalities (e.g., vision). Furthermore, the few-shot method proposed in this paper, while enhancing multilingual capabilities, also introduces a language bias centered around English.

Ethical Statement

We emphasize the importance of ethics in research involving speech data. All volunteers have signed a Voice Authorization License Agreement, granting permission for their recorded speech to be used for research purposes. The data

usage strictly complies with applicable privacy and data protection regulations. We are committed to handling all data in a responsible manner and to safeguarding the confidentiality of participants throughout the entire project.

Acknowledgements

The research in this article is supported by the National Science and Technology Major Program (Grant No. 2024ZD01NL00101), the National Science Foundation of China (U22B2059, 62276083, 62506182), the 5G Application Innovation Joint Research Institute's Project (A003), and the Major Key Project of PCL (Grant No. PCL2025A12, PCL2025A03).

References

Abouelenin, A.; Ashfaq, A.; Atkinson, A.; Awadalla, H.; Bach, N.; Bao, J.; Benhaim, A.; Cai, M.; Chaudhary, V.; Chen, C.; et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Chen, Y.; Yue, X.; Zhang, C.; Gao, X.; Tan, R. T.; and Li, H. 2024. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Conneau, A.; Ma, M.; Khanuja, S.; Zhang, Y.; Axelrod, V.; Dalmia, S.; Riesa, J.; Rivera, C.; and Bapna, A. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, 798–805. IEEE.

Cui, M.; Gao, P.; Liu, W.; Luan, J.; and Wang, B. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. *arXiv preprint arXiv:2502.02481*.

Du, Y.; Pan, Y.; Ma, Z.; Yang, B.; Yang, Y.; Deng, K.; Chen, X.; Xiang, Y.; Liu, M.; and Qin, B. 2025. Making LLMs Better Many-to-Many Speech-to-Text Translators with Curriculum Learning. arXiv:2409.19510.

Faisal, F.; Keshava, S.; Alam, M. M. I.; and Anastasopoulos, A. 2021. SD-QA: Spoken Dialectal Question Answering for the Real World. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3296–3315.

Ghosh, S.; Seth, A.; Kumar, S.; Tyagi, U.; Evuru, C. K. R.; S, R.; Sakshi, S.; Nieto, O.; Duraiswami, R.; and Manocha, D. 2024. CompA: Addressing the Gap in Compositional Reasoning in Audio-Language Models. In *The Twelfth International Conference on Learning Representations*.

He, Y.; Li, S.; Liu, J.; Tan, Y.; Wang, W.; Huang, H.; Bu, X.; Guo, H.; Hu, C.; Zheng, B.; Lin, Z.; Liu, X.; Sun, D.; Lin, S.; Zheng, Z.; Zhu, X.; Su, W.; and Zheng, B. 2024. Chinese SimpleQA: A Chinese Factuality Evaluation for Large Language Models. arXiv:2411.07140.

Jiang, B.; Zhu, R.; Wu, J.; Jiang, Z.; He, Y.; Gao, J.; Yu, J.; Min, R.; Wang, Y.; Yang, H.; et al. 2025. Evaluating Large Language Model with Knowledge Oriented Language Specific Simple Question Answering. *arXiv preprint arXiv:2505.16591*.

Li, J.; Cheng, X.; Zhao, W. X.; Nie, J.-Y.; and Wen, J.-R. 2023. HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models.

Li, Y.; Hu, B.; Chen, X.; Ma, L.; Xu, Y.; and Zhang, M. 2024. LMEye: An Interactive Perception Network for Large Language Models. *IEEE Transactions on Multimedia*, 26: 10952–10964.

Li, Y.; Jiang, S.; Hu, B.; Wang, L.; Zhong, W.; Luo, W.; Ma, L.; and Zhang, M. 2025. Uni-MoE: Scaling Unified Multimodal LLMs With Mixture of Experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5): 3424–3439.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252.

Liu, K.; Pan, Y.; Xiang, Y.; He, D.; Li, J.; Du, Y.; and Gao, T. 2025. ProjectEval: A Benchmark for Programming Agents Automated Evaluation on Project-Level Code Generation. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 20205–20221. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Longpre, S.; Lu, Y.; and Daiber, J. 2021. MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. *Transactions of the Association for Computational Linguistics*, 9: 1389–1406.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *Proceedings of ICML,2023*.

Rawte, V.; Sheth, A.; and Das, A. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*.

Wan, Z.; Yang, C.-H. H.; Yu, Y.; Tian, J.; Li, S.; Hu, K.; Chen, Z.; Watanabe, S.; Cheng, F.; Chu, C.; et al. 2025. SpeechIQ: Speech-Agentic Intelligence Quotient Across Cognitive Levels in Voice Understanding by Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 30381–30398.

Wei, J.; Karina, N.; Chung, H. W.; Jiao, Y. J.; Papay, S.; Glaese, A.; Schulman, J.; and Fedus, W. 2024. Measuring short-form factuality in large language models.

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Yu, J.; Wang, Y.; Zhong, Q.; Luo, G.; Mao, Y.; Sun, K.; Feng, W.; Xu, W.; Cao, S.; Zeng, K.; et al. 2021. MOOCCubeX: a large knowledge-centered repository for adaptive

learning in MOOCs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 4643–4652.

Zhang, Y.; Liu, X.; Zhou, R.; Chen, Q.; Fei, H.; Lu, W.; and Qin, L. 2025. CCHall: A Novel Benchmark for Joint Cross-Lingual and Cross-Modal Hallucinations Detection in Large Language Models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 30728–30749. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.