MCGA: A Multi-task Classical Chinese Literary Genre Audio Corpus

Anonymous ACL submission

Abstract

With the rapid advancement of Multimodal Large Language Models (MLLMs), their potential has garnered significant attention in Chinese Classical Studies (CCS). While existing research has primarily focused on text and visual modalities, the audio corpus within this domain remains largely underexplored. To bridge this gap, we propose the **Multi-task Classical Chinese Literary Genre Audio Corpus (MCGA)**. It encompasses a diverse range of literary genres across six specialized audio tasks: Automatic Speech Recognition (ASR), Speech-to-Text Translation (S2TT), Speech Emotion Captioning (SEC), Spoken Question Answering (SQA), Speech Understanding, and Speech Reasoning. Through the evaluation of ten MLLMs, our experimental results demonstrate that current models still face substantial challenges when processed on the MCGA test set. Furthermore, we introduce an evaluation metric for SEC and a metric to measure the consistency between the speech and text capabilities of MLLMs. We will release MCGA and our code to the public to facilitate the development of MLLMs with more robust multidimensional audio capabilities in CCS.

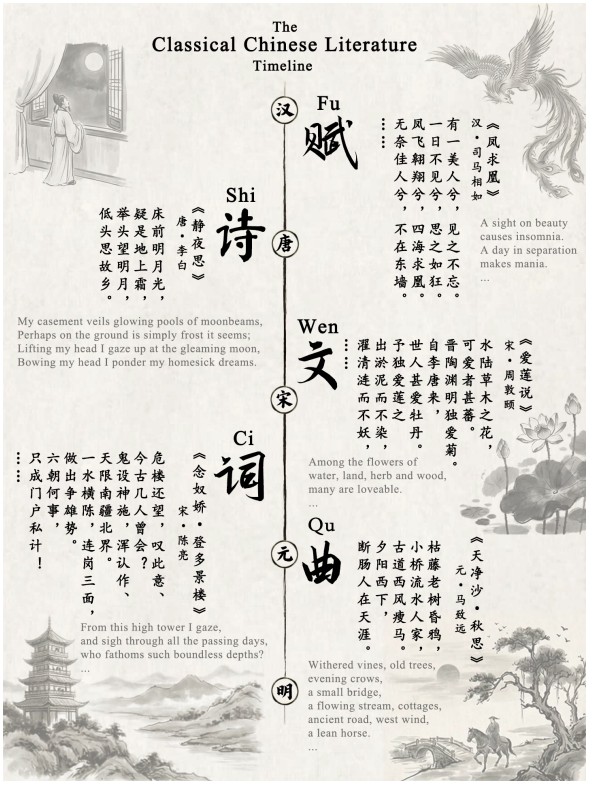


Figure 1: **Timeline of the Golden Age for Classical Chinese Literary Genres**: *Fu* (Rhapsody), *Shi* (Poetry), *Wen* (Prose), *Ci* (Lyric), and *Qu* (Song).

1 Introduction

The development of Multimodal Large Language Models (MLLMs) has significantly advanced Chinese Classical Studies (CCS). These models support multimodal inputs, providing powerful capabilities for interpreting ancient texts, which in turn enhances cultural preservation and international communication (Zhang et al., 2025a). However, while most existing research focuses on textual (Cao et al., 2024) or visual (Liu et al., 2025b) modalities, the auditory dimension of CCS remains largely unexplored. This gap stems from a lack of high-quality, domain-specific audio corpora, thereby constraining the potential for an omni-modal understanding of CCS.

To bridge this critical gap, we introduce the **Multi-task Classical Chinese Literary Genre Audio Corpus (MCGA)**, a comprehensive resource designed to catalyze audio-centric research in CCS. As illustrated in Figure 1, MCGA encompasses five primary literary genres: *Fu*, *Shi*, *Wen*, *Ci*, and *Qu*. The corpus consists of 22,000 audio samples, totaling 119 hours of recorded content. To ensure cultural and linguistic authenticity, the data were recorded by native speakers in standard Mandarin Chinese. Crucially, all audio samples include explicit copyright transfers, thereby resolving long-standing Intellectual Property Rights (IPR) challenges in open-source datasets.

Figure 2: **Examples from the MCGA Corpus.** The corpus covers six core speech tasks (ASR, S2TT, SEC, SQA, SU, SR). Leveraging its parallel speech-text data, it also supports four text tasks: Machine Translation (MT), Question Answering (QA), Language Understanding (LU), and Language Reasoning (LR).

The MCGA corpus offers two primary advantages: (1) **Task Diversity**: As illustrated in Figure 2, the corpus supports 6 diverse speech-centric tasks, including Automatic Speech Recognition (ASR), Speech-to-Text Translation (S2TT), Speech Emotion Captioning (SEC), Spoken Question Answering (SQA), Speech Understanding (SU), and Speech Reasoning (SR), alongside four integrated text tasks. (2) **Literary Genre Diversity**: It encompasses 5 major literary genres spanning 11 historical periods, forming a total of 37 distinct period-genre categories and covering a comprehensive collection of 4,497 literary works.

We evaluated 10 representative MLLMs, including 2 closed-source and 8 open-source models. Experimental results indicate that current MLLMs still have significant room for improvement in the CCS field. Notably, even the top-performing model, Qwen3-Omni (Xu et al., 2025b), scored below 60 on complex tasks such as SEC. Besides, we introduce a novel evaluation metric tailored for literary SEC, along with a Cross-Modal Consistency (CMC) metric to quantify the alignment between a model's auditory and textual reasoning. Furthermore, the substantial performance improvements observed through LoRA (Hu et al., 2022) training validate the robustness and high quality of the MCGA training set.

Our primary contributions are as follows:

- **MCGA Corpus:** We present MCGA, the first large-scale (119 hours), open-source, and fully copyrighted audio corpus dedicated to Classical Chinese literature. This resource effectively bridges the gap in high-quality audio datasets for this domain.
- **Evaluation Framework:** We establish a comprehensive evaluation framework centered on MCGA, comprising 6 multifaceted tasks: ASR, S2TT, SEC, SQA, SU, and SR. This enables a rigorous investigation into the capabilities of MLLMs.
- **Evaluation Metrics:** We introduce 2 novel evaluation metrics: a domain-specific metric tailored for **literary SEC**, and a **Cross-Modal Consistency (CMC)** metric designed to assess the alignment between auditory and textual representations.
- **Empirical Analysis:** We evaluate 10 MLLMs to identify performance bottlenecks in the classical Chinese literature domain. Besides, we demonstrate MCGA's high utility as a training resource, where fine-tuning yields substantial performance breakthroughs.

Dataset (Text† / Image✣ / Audio✧)	**Modality**	**Domain**	**Scale**	**License**	**Copyright**
ACLUE (Zhang and Li, 2023)	†	CCS	4,967	CC BY-NC-4.0	
CCLUE (Wang et al., 2023)	†	CCS	36,319	Apache-2.0	
WYWEB (Zhou et al., 2023)	†	CCS	69,700	-	
WenMind (Cao et al., 2024)	†	CCS	4,875	CC BY-NC-SA-4.0	
TianWen (Pei et al., 2025)	†	CCS	4,000	MIT	
CII-Bench (Zhang et al., 2025a)	† / ✣	General	698	Apache-2.0	
FoodieQA (Li et al., 2024)	† / ✣	Food	389	CC BY-NC-ND-4.0	✣
Oracle-Bench (Qiao et al., 2025)	† / ✣	CCS	2,834	-	
Paint4Poem (Li et al., 2021)	† / ✣	CCS	93,153	Github	
MCS-Bench (Liu et al., 2025b)	† / ✣	CCS	6,500	CC BY-NC-SA-4.0	
MCGA (ours)	† / ✧	CCS	22,000	CC BY-NC-SA-4.0	✧

Table 1: **Comparison of MCGA with existing Chinese cultural datasets**. MCGA is the first large-scale, fully copyrighted classical Chinese literary audio corpus for MLLMs (119 hours). All recordings are sourced directly from original creators with full copyright transfer, highlighting our commitment to Intellectual Property Rights (IPR) protection in Chinese Classical Studies (CCS) research.

2 Related Works

2.1 Chinese Cultural Datasets

The landscape of Chinese cultural evaluation spans many domains. ACLUE (Zhang and Li, 2023) and WYWEB (Zhou et al., 2023) establish large-scale benchmarks for Classical Chinese and ancient literature, focusing on linguistic understanding. Complementarily, CCLUE (Wang et al., 2023) rethinks cultural evaluation across broader contexts. In the multimodal sphere, FoodieQA (Li et al., 2024) and CII-Bench (Zhang et al., 2025b) probe culinary arts and figurative reasoning, respectively, highlighting a shift toward assessing complex cultural heritage and everyday traditions.

2.2 Chinese Classical Studies Datasets

Recent benchmarks deepen the evaluation of Chinese classical heritage through diverse methodologies. WenMind (Cao et al., 2024) assesses deep cultural cognition and mentalities, while TianWen (Pei et al., 2025) provides specialized assessment for traditional scriptures and historical knowledge. Advancing into multimodality, Oracle-Bench (Qiao et al., 2025) evaluates ancient script deciphering, whereas Paint4Poem (Li et al., 2021) bridges classical poetry with visual synthesis. MCS-Bench (Liu et al., 2025b) offers a framework for multimodal classical studies. However, few of these benchmarks or datasets contain the parallel speech of the classical Chinese literature.

3 MCGA Corpus

3.1 Overview

We introduce MCGA, a comprehensive corpus designed to promote audio-centric research in CCS. This section briefly outlines the construction, the human recording process, the subsequent quality control and the statistics of MCGA.

3.2 Data Construction

Data Collection and Preprocessing. Classical Chinese literature and corresponding Pinyin were sourced from the web. All works are in the public domain (created over 150 years ago). Following rigorous cleaning, texts were segmented by sentence boundaries and character counts to limit recording lengths to under 30 seconds.

Text Data Construction. Subsequently, we leverage DeepSeek-V3.2 (Liu et al., 2024) to generate question-answer pairs for the text of each clip, with access to the full literary context. This process covers a variety of speech-related tasks, including S2TT, SEC, SU, and SR.

Text Data Verification. The generated question–answer pairs are subjected to trio validation using DeepSeek-V3.2, GPT-5-mini (OpenAI, 2025), and Gemini-3-Flash (Team and DeepMind, 2025), through which pairs that fail to pass the verification are preliminarily filtered out. The test and validation sets underwent human verification to ensure data quality.

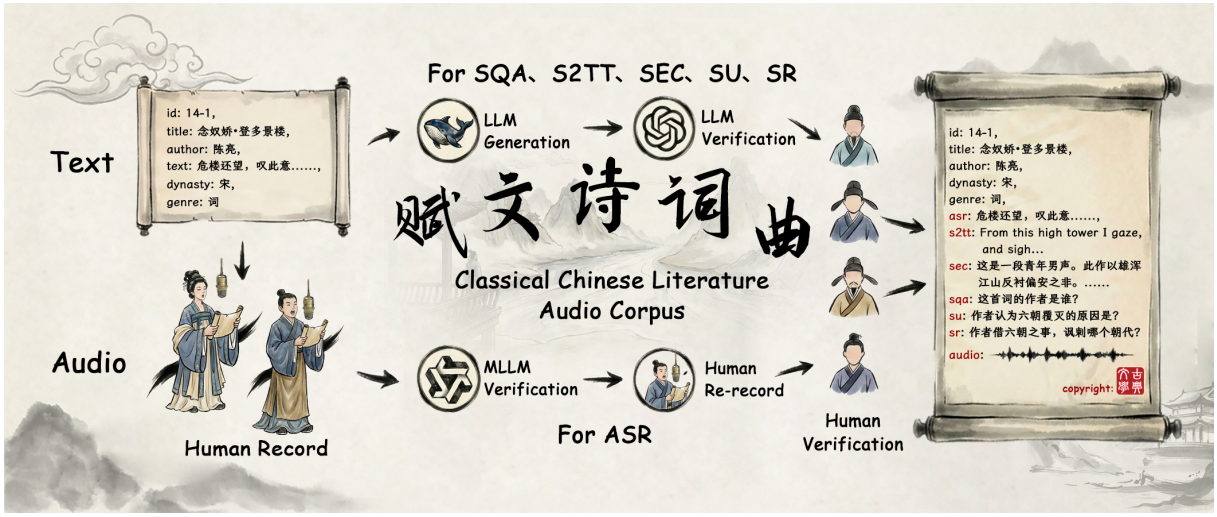


Figure 3: **MCGA Corpus Construction.** Initially comprising only metadata such as titles, authors, and texts, the MCGA corpus is expanded through human recording, LLM generation, and rigorous verification. Then, it supports six speech tasks: ASR, S2T, SEC, SQA, SU, and SR. We provide a detailed example of the SEC task in Figure 4.

3.3 Human Recording

Volunteer Demographics. We recruited 28 native speakers (13 males and 15 females, aged 18–40) to record the texts via a dedicated private website. All participants have good educational backgrounds, half of whom are Chinese majors.

Recording Protocol. We explicitly stated the recording guidelines to the volunteers, as follows:

Volunteer Guidelines

- use a tone that matches the emotion of the text.
- can access the pinyin for all Chinese characters.
- ensure the recording environment is quiet.
- each clip must be read in 30 seconds.
- each clip is read by at least 1 male and 1 female.
- clips from the same work are sent to the same person.

3.4 Audio Quality Check

MLLM Verification. We employed a dual-stage speech recognition verification process using Qwen and Whisper (Radford et al., 2023) models to identify samples with significant errors, which were subsequently re-recorded by the volunteers.

Human Verification. We recruited 6 data quality inspection volunteers to verify the validation and test sets. The inspectors were instructed to score the samples. Low quality samples (pronunciation error or presence of background noise) were removed from the sets.

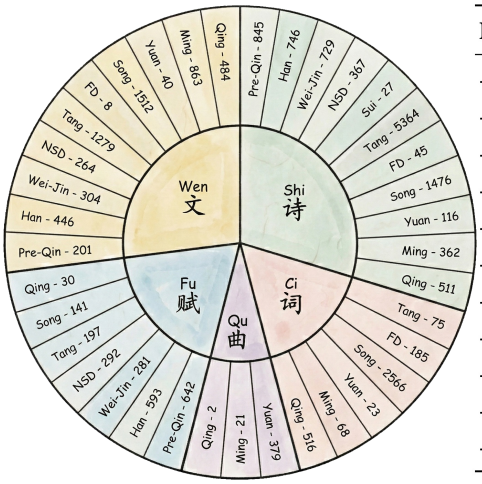
Both recording volunteers and quality inspection volunteers signed labor agreements and were compensated with reasonable remuneration.

Figure 4: **Case for SEC Task.**

3.5 Case Study for SEC

SEC Task. To capture the emotional and artistic nuances of classical literature, we present a case study in Figure 4. For SEC, the MLLM must sequentially generate three components:

1. **Persona Profiles:** Analysis of the speaker, including age and gender.
2. **Overall Sentiment Analysis:** A summary of the general emotional tone and attitude.
3. **Paired Transcription & Emotion:** A sentence-by-sentence decomposition in the format: “Transcription | Emotion.”



Period	Shi	Ci	Qu	Fu	Wen
- Pre-Qin	845	0	0	642	201
- Han	746	0	0	593	446
- Wei-Jin	729	0	0	281	304
- NSD	367	0	0	292	264
- Sui	27	0	0	0	0
- Tang	5,364	75	0	197	1,279
- FD	45	185	0	0	8
- Song	1,476	2,566	0	141	1,512
- Yuan	116	23	379	0	40
- Ming	362	68	21	0	863
- Qing	511	516	2	30	484

Statistics	22000 / 119.0h
Task	Train / Valid / Test
- ASR	20,000 / 1,000 / 1,000
- S2TT	15,163 / 1,000 / 1,000
- SEC	19,343 / 1,000 / 1,000
- SQA	20,000 / 1,000 / 1,000
- SU	17,640 / 1,000 / 1,000
- SR	17,730 / 1,000 / 1,000
Data Length	Char / Time (s)
- Maximum	158 / 30.0
- Minimum	16 / 3.5
- Average	54.7 / 19.5

Figure 5: **Corpus Statistics.** It comprises 22,000 filtered human-recorded speech samples (totaling 119 hours) and supports 6 downstream tasks. Sample counts for S2TT, SEC, SU, and SR are lower due to the removal of invalid QA pairs. (NSD: the Northern and Southern Dynasties; FD: the Five Dynasties)

3.6 Dataset Statistics

Figure 5 shows the statistics of MCGA. It spans 5 genres across 11 historical periods, resulting in 37 unique period–genre categories.

Tang Shi has the most samples, followed by Song Ci. This is because the Tang and Song dynasties were the two peak periods of classical Chinese literature. Shi was the most popular genre in the Tang dynasty, while Ci was in the Song dynasty.

The corpus comprises 22,000 filtered human-recorded speech samples (totaling 119 hours) and supports 6 downstream tasks: ASR, S2TT, SEC, SQA, SU, and SR. The longest audio sample is 30 seconds, the shortest is 3.5 seconds, and the average duration is 19.5 seconds.

It should be noted that sample counts for S2TT, SEC, SU, and SR are lower due to the removal of invalid QA pairs. Also, the validation or test sets for the six tasks are not parallel.

Task	Metric	Details
ASR	CER ↓	Text normalization Following Morris et al. (2004)
S2TT	LLM-B ↑	LLM Evaluation Following Chen et al. (2025)
SEC	LLM-C ↑	LLM Evaluation Proposed in this work
SQA	F1 ↑	Open-ended Factuality Evaluation
SU	Accuracy ↑	Multiple-choice questions Options derived from the speech
SR	Accuracy ↑	Multiple-choice questions External knowledge reasoning

Table 2: **Metric Details.**

4 Experiments

4.1 Experiment Setting

Baseline MLLMs. We evaluate 2 closed-source MLLMs (GPT-4o-mini-Audio (OpenAI, 2023) and Gemini-3-Flash (Team et al., 2025)) and 8 open-source MLLMs: the Qwen series (Chu et al., 2024; Xu et al., 2025a,b), the Voxtral series (Liu et al., 2025a), Phi-4-Multimodal-Instruct (Abouelenin et al., 2025), MiDashengLM (Dinkel et al., 2025), and Step-Audio-2-mini (Wu et al., 2025).

Training Details. We fine-tuned Qwen2.5-Omni-7B using the ms-swift framework[1] with LoRA ($r = 8, \alpha = 32$) (Hu et al., 2021). The model was trained for 3 epochs on 4 A100 GPUs using the AdamW optimizer with a learning rate of 1×10^{-4}, a per-device batch size of 8, and a gradient accumulation of 4.

Evaluation Metrics. As shown in Table 4, we evaluate MLLMs across six tasks. All open-source models are deployed using the vLLM framework[2] (Kwon et al., 2023), with inference performed via API requests at a temperature of 0. To provide a more intuitive performance metric, we normalize the S2TT and SEC results to a 100-point scale. Specifically, the ASR task is evaluated using the Character Error Rate (CER)[3], while the S2TT and SEC tasks are scored by the deepseek-chat API (Guo et al., 2025). For SQA, we report the F1 score, and for SU and SR, we report Accuracy.

[1] https://github.com/modelscope/ms-swift
[2] https://github.com/vllm-project/vllm
[3] https://github.com/jitsi/jiwer

Model	ASR CER ↓	S2TT LLM-B ↑	SEC LLM-C ↑	SQA F1 ↑	SU Acc ↑	SR Acc ↑
Closed-source Models						
GPT-4o-mini-Audio (OpenAI, 2023)	20.5	42.2	6.0	31.2	75.0	70.4
Gemini-3-Flash (Team et al., 2025)	7.0	**74.5**	54.0	48.7	86.6	**83.7**
Open-source Models						
Phi-4-Multimodal-Instruct (Abouelenin et al., 2025)	58.2	26.2	12.7	24.5	51.2	54.5
Voxtral-Mini (Liu et al., 2025a)	28.0	24.7	14.8	12.4	59.6	62.5
Voxtral-Small (Liu et al., 2025a)	29.3	33.2	14.6	28.3	72.9	71.8
MiDashengLM (Dinkel et al., 2025)	11.7	43.9	22.6	22.5	72.2	75.6
Step-Audio-2-mini (Wu et al., 2025)	10.1	43.0	39.9	45.3	80.7	79.9
Qwen2-Audio-7B-Instruct (Chu et al., 2024)	18.9	30.6	24.1	25.1	71.9	64.7
Qwen2.5-Omni-7B (Xu et al., 2025a)	10.5	50.5	36.8	43.2	81.3	79.5
Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b)	**4.4**	70.0	**59.5**	**51.8**	**87.1**	83.2

Table 3: **Performance comparison of different models on the MCGA test set.** Complete results for LLM-B and LLM-C are shown in Table 6 and Table 7.

4.2 Main Results

We present a comprehensive evaluation of ten MLLMs across six audio tasks. By analyzing the interplay between model performance and task difficulty, we derive the following key observations:

Closed-source vs. Open-source Models. In Table 3, Qwen3-Omni demonstrates superior performance on the MCGA test set for Chinese understanding and generation tasks, specifically in **ASR** (4.4 CER ↓), **SEC** (59.5 LLM-C ↑), **SQA** (51.8 F1 ↑), and **SU** (87.1 Acc ↑). Conversely, closed-source models such as Gemini-3-Flash maintain a competitive edge in English generation and Chinese reasoning tasks, leading in metrics such as **S2TT** (74.5 LLM-B ↑) and **SR** (83.7 Acc ↑). Overall, the open-source models have achieved a competitive level of performance compared with the closed-source ones.

Comparison across Different Tasks. As shown in Figure 6, existing MLLMs demonstrate their strongest performance in Chinese classical literature **ASR** tasks. This is followed by **SU** and **SR** in multiple-choice formats, where models achieve relatively robust results. Regarding the **S2TT** task, overall performance is acceptable but remains to be enhanced. In contrast, performance on **SEC** is notably poor, indicating a critical need for enhanced affective computing capabilities. Finally, for open-ended **SQA**, F1 scores remain low, suggesting that audio "hallucination" issues (Du et al., 2025) have yet to be effectively resolved.

4.3 Further Analysis

4.3.1 Analysis of ASR Task

Performance Disparity Across Genres. Table 4 shows the MLLMs' performance which varies significantly by genre. Qwen3-Omni achieves state-of-the-art results on MCGA, maintaining the lowest CER across all categories, particularly in *Ci* (2.9). A consistent trend across various models is that *Ci* achieves lower CER, while *Fu* consistently poses the greatest difficulty. This difficulty stems from *Fu*'s ornate rhetoric, frequent classical allusions, and high density of modal particles.

Models	Shi	Ci	Qu	Fu	Wen
GPT-4o-mini-Audio	22.9	20.1	18.4	22.8	18.3
Gemini-3-Flash	6.1	6.8	7.7	8.4	6.1
Phi-4-Multimodal-Instruct	58.5	61.3	62.5	60.0	50.7
Voxtral-Mini	27.4	24.7	27.8	32.4	27.1
Voxtral-Small	30.5	24.7	29.6	32.7	28.7
MiDashengLM	12.7	10.1	9.4	15.7	10.0
Step-Audio-2-mini	9.0	6.8	7.5	15.1	10.8
Qwen2-Audio-7B-Instruct	19.1	16.7	15.9	23.1	18.7
Qwen3-Omni-30B-A3B-Instruct	**3.8**	**2.9**	**3.8**	**6.4**	**4.6**
Qwen2.5-Omni-7B	11.6	7.8	8.8	15.1	8.6
Qwen-Omni-MCGA	**2.8**	**2.9**	7.7	**5.2**	**4.2**

Table 4: **CER Scores Across Different Genres.** The test set contains 1,000 samples (200 per genre). Underline indicates the best-performing genre for each individual model. Qwen-Omni-MCGA is a LoRA-based adaptation of Qwen2.5-Omni-7B. It achieves state-of-the-art results on all genres except for *Qu*.

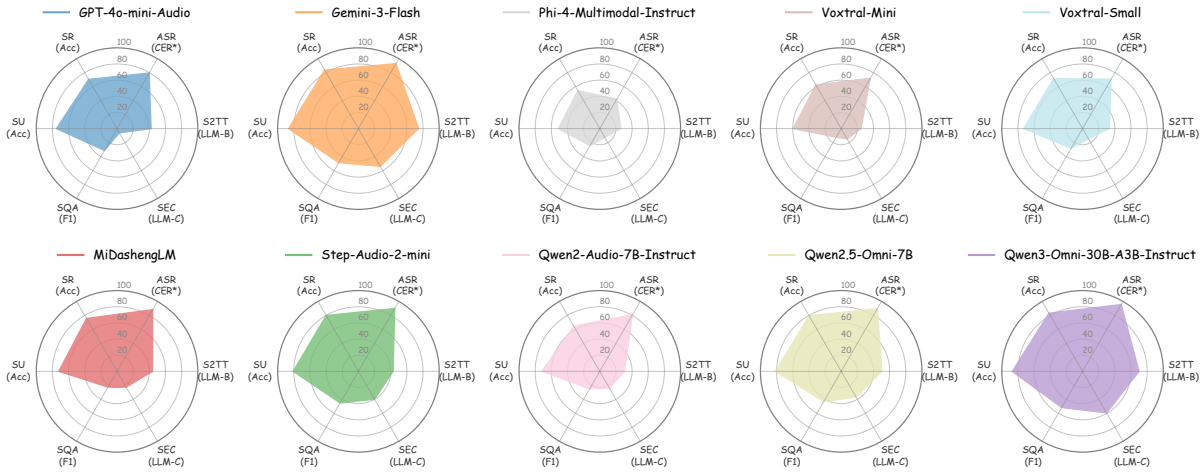


Figure 6: **Comparison across Different Tasks.** Existing MLLMs exhibit robust performance in **ASR**, **SU**, and **SR** tasks, but they still encounter challenges regarding the beauty of translation in **S2TT**, affective modeling in **SEC**, and hallucination issues in open-ended **SQA**. CER* refers to $(1 - \text{CER}\%)$.

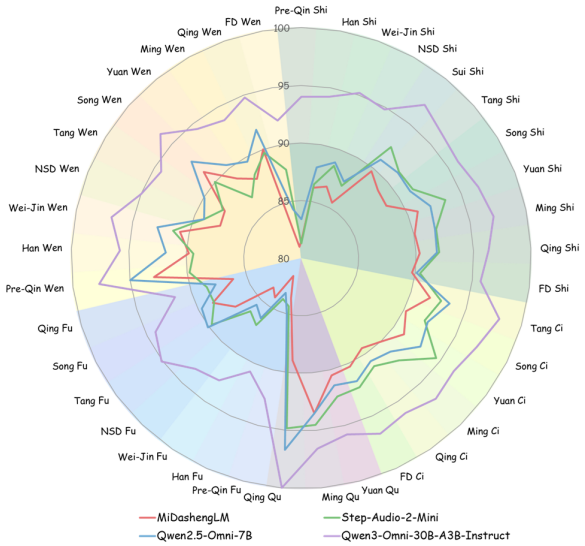


Figure 7: **CER* Across Dynasties and Genres.** CER* refers to $(1 - \text{CER}\%)$

ASR Quality. Table 5 reveals a 0.2 CER gap (Qwen3-Omni) between human-verified valid/test sets and the train set, confirming high data consistency. Residual errors primarily stem from uncommon characters and phonetic loanwords (*tongjiazi*) in Classical Chinese. Figure 7 shows the CER distribution across dynasties and genres.

Models	Train	Valid	Test	Avg.
MiDashengLM	11.7	11.2	11.7	11.7
Step-Audio-2-mini	10.5	10.1	10.1	10.5
Qwen2.5-Omni-7B	9.8	9.5	10.5	9.8
Qwen3-Omni-30B-A3B-Instruct	4.6	4.4	4.4	4.6

Table 5: **CER Scores for Quality Check.** The train, valid, and test sets show high data consistency.

4.3.2 Analysis of S2TT Task

Beauty Evaluation of Translation. As illustrated in Table 6, we evaluate the translation quality across four dimensions: Beauty of Form (LLM-BF), Beauty of Meaning (LLM-BM), Beauty of Sound (LLM-BS), and their average score (LLM-B). The closed-source model **Gemini-3-Flash** achieves the highest performance across all metrics, reaching a peak average score of 74.0 (LLM-B ↑).

S2TT Quality. Additionally, we provide high-quality ground-truth translation candidates. The LLM-B score of 79.2 (4.0) is constrained by the 1–5 evaluation scale, as the DeepSeek API evaluation model typically assigns moderate scores and rarely grants a perfect score of 5. To provide a more intuitive performance metric, we normalize these raw API scores to a 100-point scale.

Models	LLM-BF	LLM-BM	LLM-BS	LLM-B
Ground Truth	79.8	80.3	77.4	79.2
GPT-4o-mini-Audio	40.7	41.7	44.2	42.2
Gemini-3-Flash	**72.4**	**74.5**	**75.2**	**74.0**
Phi-4-Multimodal-Instruct	26.0	26.2	27.2	26.5
Voxtral-Mini	24.7	24.7	23.9	24.4
Voxtral-Small	33.3	33.2	32.7	33.1
MiDashengLM	43.0	43.9	40.1	42.4
Step-Audio-2-mini	42.8	43.0	39.8	41.9
Qwen2-Audio-7B-Instruct	29.5	30.6	30.7	30.3
Qwen2.5-Omni-7B	50.0	50.5	45.3	48.6
Qwen3-Omni-30B-A3B-Instruct	68.7	70.0	66.5	68.4

Table 6: **Beauty Evaluation of Translation.** Following Chen et al. (2025), we employ Beauty of Form (BF), Beauty of Meaning (BM) and Beauty of Sound (BS) as evaluation metrics. LLM-B denotes the average score.

4.3.3 Analysis of SEC Task

SEC Evaluation. We design an LLM-based penalty evaluation mechanism based on reference answers. The mechanism consists of the following three metrics:

- **Persona Recognition (SEC-P, 0–2):** Measures the capability to extract identity features such as age and gender. Starting from an initial score of 2, 1 point is deducted for each attribute error.
- **Global Emotional Tone (SEC-G, 0–3):** Evaluates the overall emotional atmosphere based on the richness and accuracy of the descriptions. A score of 0 is assigned if the emotional category or context is misidentified.
- **Sentence-level Emotion Tracking (SEC-S, 0–5):** Evaluates sentence-by-sentence transcription and analysis. 1 point is deducted for each error in emotional portrayal. If the transcription is entirely unrelated (hallucination), a score of 0 is recorded.

Open-source vs. Closed-source MLLMs. As shown in Table 7, Qwen3-Omni outperforms other models across all SEC metrics. This superior performance is attributed to its deep understanding of Chinese cultural nuances and its robust transcription capabilities. It is followed by Gemini-3-Flash, which maintains competitive results.

In contrast, GPT-4o-mini-Audio exhibits poor performance. This is primarily because its stringent safety protocols frequently trigger refusals when tasked with persona-based or emotional analysis.

Models	SEC-P	SEC-G	SEC-S	LLM-C
Ground Truth	20.0	30.0	50.0	100.0
GPT-4o-mini-Audio	1.5	3.2	1.2	6.0
Gemini-3-Flash	13.4	16.9	23.6	54.0
Phi-4-Multimodal-Instruct	4.6	7.7	0.4	12.7
Voxtral-Mini	4.9	9.1	0.8	14.8
Voxtral-Small	1.7	10.8	2.1	14.6
MiDashengLM	13.1	6.6	2.9	22.6
Step-Audio-2-mini	16.3	12.6	11.1	39.9
Qwen2-Audio-7B-Instruct	10.2	11.4	2.5	24.1
Qwen2.5-Omni-7B	14.1	13.9	8.8	36.8
Qwen3-Omni-30B-A3B-Instruct	**16.1**	**19.0**	**24.4**	**59.5**

Table 7: **LLM-based Evaluation for SEC.** (1) SEC-P (0–2) for persona identification; (2) SEC-G (0–3) for global emotional tone analysis; (3) SEC-S (0–5) for sentence-level emotion; (4) LLM-C is the sum of scores.

4.3.4 Analysis of SQA, SU, and SR Tasks

Open-ended vs. Multiple-choice QA. As shown in Table 8, a substantial performance gap exists between multiple-choice and open-ended formats. MLLMs struggle significantly more with open-ended questions, such as identifying authors or titles, compared to complex reasoning tasks that provide candidate options. For instance, Gemini-3-Flash scored 86.6 in SU and 83.7 in SR but drops to 48.7 in SQA. This gap indicates that MLLMs suffer from severe hallucinations in open-ended factual QA, despite their strong reasoning.

Cross-modal Consistency. To evaluate how reliably MLLMs maintain consistency across different input modalities, we define the Cross-modal Consistency (CMC) metric as:

$$\mathrm{CMC} = \frac{1}{3}\left(\frac{\mathrm{SQA}}{\mathrm{QA}} + \frac{\mathrm{SU}}{\mathrm{LU}} + \frac{\mathrm{SR}}{\mathrm{LR}}\right) \times 100 \quad (1)$$

As shown in Table 8, SQA, SU, and SR represent performance on speech-based tasks, while the denominators QA, LU (Language Understanding), and LR (Language Reasoning) serve as the text-only upper-bound references. Step-Audio-2-mini achieved the highest CMC score among all evaluated MLLMs.

Models	SQA	SU	SR	QA	LU	LR	CMC
Gemini-3-Flash	48.7	86.6	**83.7**	**66.0**	**94.6**	**91.5**	85.6
Phi-4-Multimodal-Instruct	24.5	51.2	54.5	25.1	69.5	61.0	86.9
Voxtral-Mini	12.4	59.6	62.5	13.6	77.5	69.5	86.0
Voxtral-Small	28.3	72.9	71.8	39.6	89.9	83.6	79.5
MiDashengLM	22.5	72.2	75.6	41.7	89.6	85.0	74.5
Step-Audio-2-mini	45.3	80.7	79.9	51.3	89.8	85.8	**90.4**
Qwen2-Audio-7B-Instruct	25.1	71.9	64.7	35.6	79.6	71.1	83.9
Qwen2.5-Omni-7B	43.2	81.3	79.5	53.6	91.3	84.9	87.8
Qwen3-Omni-30B-A3B-Instruct	**51.8**	**87.1**	83.2	60.4	93.6	91.0	90.1

Table 8: **Cross-modal Consistency.** CMC quantifies the performance gap between audio and textual modalities.

5 Conclusion

This paper introduces **MCGA**, the first large-scale, fully copyrighted audio corpus for classical Chinese literature, featuring six speech-language tasks. We develop an evaluation metric for literary SEC and a metric to assess cross-modal consistency. Our systematic evaluation of 10 MLLMs shows that the Qwen-series models demonstrated superior proficiency in understanding CCS.

6 Limitations

Although MCGA incorporates audio-text multimodal data across six distinct tasks, several limitations persist. First, copyright constraints preclude the inclusion of real-world image samples that are precisely aligned with both textual and auditory modalities. Second, the *Qu* genre emerged significantly later than *Shi*, *Ci*, *Wen*, and *Fu*. Due to the relatively short-lived nature of the Yuan Dynasty, the volume of extant works is considerably limited, leading to a lower representation of *Qu* within the corpus.

7 Ethical considerations

We emphasize that ethical standards are of paramount importance in research involving human audio data. All audio data used in this study were recorded by human volunteers who contacted the authors directly. The volunteers were fairly compensated for their contributions and have signed a Voice Authorization License Agreement, explicitly granting permission for their recorded speech to be used for research purposes. Data handling and usage strictly comply with all applicable privacy and data protection regulations. All audio data in the final corpus have been anonymized.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Jiahuan Cao, Yang Liu, Yongxin Shi, and 1 others. 2024. Wenmind: A comprehensive benchmark for chinese classical literature and language arts. In *NeurIPS 2024 Datasets and Benchmarks Track*.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2025. Benchmarking LLMs for translating classical Chinese poetry: Evaluating adequacy, fluency, and elegance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33007–33024, Suzhou, China. Association for Computational Linguistics.

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

Heinrich Dinkel, Gang Li, Jizhong Liu, Jian Luan, Yadong Niu, Xingwei Sun, Tianzi Wang, Qiyang Xiao, Junbo Zhang, and Jiahao Zhou. 2025. Midashenglm: Efficient audio understanding with general audio captions. *arXiv preprint arXiv:2508.03983*.

Yexing Du, Kaiyuan Liu, Youcheng Pan, Zheng Chu, Bo Yang, Xiaocheng Feng, Ming Liu, and Yang Xiang. 2025. Ccfqa: A benchmark for cross-lingual and cross-modal speech and text factuality evaluation. *Preprint*, arXiv:2508.07295.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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

Dan Li, Shuai Wang, Jie Zou, and 1 others. 2021. Paint4poem: A dataset for artistic visualization of classical chinese poems. *arXiv preprint arXiv:2109.11682*.

Wenyan Li, Crystina Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, Daniel Hershcovich, and Desmond Elliott. 2024. FoodieQA: A multimodal dataset for fine-grained understanding of Chinese food culture. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19077–19095, Miami, Florida, USA. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv e-prints*, pages arXiv–2412.

Alexander H Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, and 1 others. 2025a. Voxtral. *arXiv preprint arXiv:2507.13264*.

Yang Liu, Jiahuan Cao, Hiuyi Cheng, Yongxin Shi, Kai Ding, and Lianwen Jin. 2025b. Mcs-bench: A comprehensive benchmark for evaluating multimodal large language models in chinese classical studies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10435–10492.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

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

OpenAI. 2025. GPT-5 System Card. Technical report, OpenAI. Technical Report.

Zhenwu Pei, Rongbo Chen, Xuefeng Bai, Kehai Chen, Yingjie Zhu, Andong Chen, and Min Zhang. 2025. Tianwen: A comprehensive benchmark for evaluating llms in chinese classical poetry understanding and reasoning. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 516–528. Springer.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Jiapeng Wang, YiFan Zhang, Zhuoma GongQue, Chong Sun, Yida Xu, Yadong Xue, Ye Tian, Zhimin Bao, Lan Yang, Chen Li, and Honggang Zhang. 2025. V-oracle: Making progressive reasoning in deciphering oracle bones for you and me. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20124–20150, Vienna, Austria. 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*.

Gemini Team and Google DeepMind. 2025. Gemini 3 Flash Model Card. Technical report, Google. Technical Report.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Yuxuan Wang, Jack Wang, Dongyan Zhao, and Zilong Zheng. 2023. Rethinking dictionaries and glyphs for Chinese language pre-training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1089–1101, Toronto, Canada. Association for Computational Linguistics.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, and 90 others. 2025. Step-audio 2 technical report. *Preprint*, arXiv:2507.16632.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, and 1 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Chenhao Zhang, Xi Feng, Yuelin Bai, Xeron Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, Xingwei Qu, Yifei Zhang, Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang, Min Yang, Wenhao Huang, Chenghua Lin, and 2 others. 2025a. Can MLLMs understand the deep implication behind Chinese images? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14402, Vienna, Austria. Association for Computational Linguistics.

Chenhao Zhang, Xi Feng, Yuelin Bai, Xeron Du, Jinchang Hou, Kaixin Deng, Guangzeng Han, Qinrui Li, Bingli Wang, Jiaheng Liu, Xingwei Qu, Yifei Zhang, Qixuan Zhao, Yiming Liang, Ziqiang Liu, Feiteng Fang, Min Yang, Wenhao Huang, Chenghua Lin, and 2 others. 2025b. Can MLLMs understand the deep implication behind Chinese images? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14402, Vienna, Austria. Association for Computational Linguistics.

Yixuan Zhang and Haonan Li. 2023. Can large language model comprehend Ancient Chinese? a preliminary test on ACLUE. In *Proceedings of the Ancient Language Processing Workshop*, pages 80–87, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Bo Zhou, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. 2023. WYWEB: A NLP evaluation benchmark for classical Chinese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3294–3319, Toronto, Canada. Association for Computational Linguistics.