Improving Scientific Formula Verbalization in Large Speech Language Models for Accessible Learning

Xueyi Li[1], **Tianqiao Liu**[2], **Zitao Liu**[1], **Teng Guo**[1*] and **Yongdong Wu**[1]
[1]Guangdong Institute of Smart Education, Jinan University
[2]TAL Education Group
lixueyi@stu2021.jnu.edu.cn, liutianqiao1@tal.com,
{liuzitao, tengguo, wuyongdong}@jnu.edu.cn

Abstract

Online learning systems provide accessible learning opportunities for blind or low-vision students. To support access to complex scientific materials, the speech models used in these systems need to deliver accurate scientific formula verbalization. While recent large speech language models (LSLMs) provide remarkable low-latency streaming capabilities, their potential for scientific formula verbalization remains underexplored. In this paper, we propose Formula-Speech, the first end-to-end LSLM designed for scientific formula verbalization. Specifically, we construct two high-quality scientific formula datasets with educational experts to align speech models with scientific formula verbalization patterns. We then adopt a lightweight and effective two-stage training framework, combining supervised fine-tuning for basic formula-to-speech alignment with reinforcement learning guided by a custom reward function to optimize for human-preferred verbalization. Experimental results show that our model significantly improves the verbalization performance of LSLMs and achieves state-of-the-art results across multiple scientific domains. Our code and datasets are available at https://github.com/ai4ed/FormulaSpeech.

1 Introduction

Accessible learning aims to provide fair educational opportunities in which students with different abilities can fully participate, and it has become an important part of online education [Hadi Mogavi *et al.*, 2023; Shi *et al.*, 2024]. Many online learning systems use AI tutors to support students, especially blind or low-vision students who rely on spoken feedback as their primary way of accessing course content [Pal Chowdhury *et al.*, 2024; Anderer *et al.*, 2025]. When courses involve scientific formulas, students depend entirely on how the formulas are read to understand the expression, which highlights the importance of accurate formula verbalization in AI tutors [Gorniak *et al.*, 2024; Abram, 2025; Chen *et al.*, 2025].

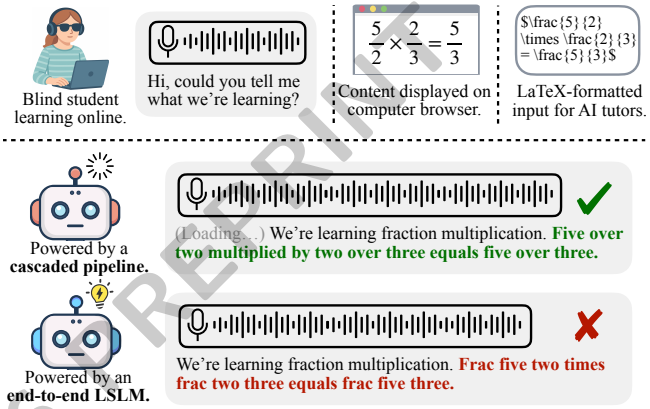

Figure 1: Example of formula verbalization for accessible learning.

Previous research on scientific formula verbalization has mainly used cascaded pipelines, which are often adopted in AI tutors. For example, Sieun Hyeon et al. propose a pipeline that integrates automatic speech recognition (ASR) models with language models to correct errors in mathematical expressions and accurately convert spoken expressions to speech output by text-to-speech (TTS) models [Hyeon *et al.*, 2025]. With the development of large language models (LLMs) [Vaswani *et al.*, 2017], recent studies have introduced end-to-end large speech language models (LSLMs) that extend the natural language understanding and representation capabilities of LLMs to the speech modality through multimodal alignment techniques [Zeng *et al.*, 2025; Liu *et al.*, 2026]. Compared to cascaded pipelines, end-to-end LSLMs have attracted increasing attention due to their low latency, capability for streaming generation and potential for industrial deployment. However, when directly applying these LSLMs to AI tutors, their performance in scientific formula verbalization remains suboptimal, with frequent mispronunciations of certain symbols (as shown in Figure 1). Achieving accurate scientific formula verbalization with LSLMs continues to pose an open research challenge.

To improve scientific formula verbalization in end-to-end LSLMs, a promising approach is to directly incorporate high-quality paired data of formula text and corresponding speech during training. Similar to large-scale LLM pretraining, aligning the textual and speech modalities in LSLMs

*Corresponding author: Teng Guo.

during pretraining also requires massive amounts of data. For example, GLM-4-Voice and Kimi-Audio are trained on over 700,000 hours of speech data [Zeng *et al.*, 2025; Ding *et al.*, 2025]. However, large-scale text-speech paired datasets specifically curated for scientific formulas are difficult to obtain, making formula-oriented large-scale pretraining impractical. This motivates a shift toward the supervised fine-tuning (SFT) stage, where task-specific capabilities can be improved with a smaller amount of high-quality paired data. As LSLMs are developed on top of LLMs, the alignment between general text and speech patterns has largely been established through pretraining, equipping these models with strong capabilities in understanding scientific formulas in textual form, but a small yet critical gap remains in aligning certain scientific symbolic expressions with corresponding speech. To bridge this gap, we construct a high-quality text-speech paired dataset for scientific formulas and design a lightweight and effective method to investigate the following research question: *How to align the text and speech modalities in LSLMs for scientific formula verbalization with limited text-speech paired data?*

In this paper, we propose *Formula-Speech*, the first end-to-end LSLM designed to enhance scientific formula verbalization by efficiently aligning symbolic text with its corresponding speech. Specifically, we construct two high-quality datasets covering three scientific domains including mathematics, physics and chemistry, in collaboration with educational experts. One dataset consists of paired formula-speech data for aligning symbolic expressions with spoken language, while the other comprises multi-turn teacher-student dialogues in a speech-to-speech format to support the learning of conversational speech features. We further introduce a lightweight and effective two-stage training framework based on low-rank adaptation (LoRA). In the first stage, we perform SFT with LoRA to help the model learn fundamental mapping patterns between scientific formulas and their spoken representations. In the second stage, we employ a reinforcement learning (RL) algorithm, i.e., group relative policy optimization (GRPO), combined with a custom designed LLM-as-a-Judge reward function. This reward function evaluates ASR-transcribed outputs and guides the model toward human-preferred verbalization. Experimental results show that the proposed Formula-Speech consistently outperforms strong baselines across all three domains and achieves state-of-the-art performance in scientific formula verbalization despite being trained on only a small amount of data.

Overall, this paper makes the following contributions:

- We propose the first end-to-end LSLM optimized for scientific formula verbalization. It adopts a lightweight and effective two-stage training framework that requires only a small amount of scientific formula data.
- We construct and release two high-quality datasets covering mathematics, physics and chemistry to facilitate future research in the community.
- We conduct comprehensive experiments across three scientific domains to evaluate the effectiveness of the proposed Formula-Speech, demonstrating state-of-the-art performance on scientific formula verbalization.

2 Related Work

2.1 Formula Verbalization

In online learning systems, AI tutors often use spoken explanations to help students access course content, especially those who cannot easily rely on visual materials. This motivates research on scientific formula verbalization, which aims to convert symbolic expressions such as LaTeX formulas into clear and standardized spoken language. Early approaches primarily treated formula verbalization as a speech processing task, leveraging traditional methods such as hidden Markov models and rule based systems [Prylipko *et al.*, 2012; Sak *et al.*, 2013]. In pursuit of better performance and robustness, recent research has shifted toward adopting cascaded pipelines, focusing predominantly on the mathematical domain. Ritchie et al. introduced a unified grammar based on finite-state transducers to jointly support ASR and TTS, aiming for consistent cross-lingual formula rendering [Ritchie *et al.*, 2019]. Dong et al. proposed a dictionary-guided generation model that significantly outperformed rule based baselines in the textual verbalization of mathematical formulas [Dong *et al.*, 2022]. Hyeon et al. developed a cascaded system that combines ASR and symbolic correction with a language model, achieving superior performance in formula transcription [Hyeon *et al.*, 2025]. Roychowdhury et al. conducted a systematic evaluation of TTS models for formula pronunciation, revealing that current systems struggle with complex mathematical structures [Roychowdhury *et al.*, 2025]. Different from previous approaches, we focus on improving scientific formula verbalization in end-to-end LSLMs across the domains of mathematics, physics and chemistry.

2.2 Large Speech Language Models

Current speech interaction in AI tutors mostly adopts cascaded pipelines, typically composed of ASR, LLMs and TTS modules. While this architecture is functional, it suffers from several limitations, including high latency and error accumulation across stages [Fang *et al.*, 2025; Xu *et al.*, 2025; Huang *et al.*, 2025]. To address these challenges, recent research has focused on extending LLMs to the speech modality, giving rise to end-to-end LSLMs. These models aim to unify speech understanding and generation within a single framework, enabling seamless and low latency speech interaction. Zeng et al. introduced an interleaved speech-text training paradigm and a supervised speech tokenizer, achieving strong performance in spoken question answering and conversational ability [Zeng *et al.*, 2025]. Li et al. further advanced the field with a multi-codebook tokenizer and independent speech head, along with a two-stage pretraining strategy to preserve LLM capabilities during speech integration [Li *et al.*, 2025]. Building on these advances, Long et al. tackled the latency bottleneck by proposing a fast interleaved token generation mechanism, achieving zero-delay speech response via efficient token prediction modules [Long *et al.*, 2025]. While these end-to-end LSLMs have demonstrated impressive progress in general conversational tasks, they largely overlook domain-specific challenges such as scientific formula verbalization, which is the primary focus of our work.

3 Problem Definition

To support accessible learning with AI tutors in online learning systems, we aim to convert a given scientific formula into spoken language that accurately verbalizes each symbolic component. Formally, we define a scientific formula $x \in \mathcal{X}$ as a sequence of n tokens, $x = \{t_1, t_2, \ldots, t_n\}$, where each t_i denotes a symbolic token and n is the formula length. Given a paired dataset $\mathcal{D} = \{(x_{(i)}, y_{(i)})\}_{i=1}^{N}$ of N examples, where $x_{(i)}$ is a formula and $y_{(i)} \in \mathcal{Y}$ is its corresponding spoken form represented as speech tokens, we train a speech language model $\mathbf{G} : \mathcal{X} \to \mathcal{Y}$ that maps input formulas x to generated speech $\hat{y} = \mathbf{G}(x)$. Unlike natural language, scientific formulas contain compact syntax, hierarchical structure and domain-specific semantics, posing unique challenges for general-purpose speech models.

4 Method

4.1 Dataset Construction

To improve scientific formula verbalization in LSLMs, high-quality text-speech paired datasets are essential for aligning symbolic representations with corresponding speech patterns. However, existing resources are limited in both domain coverage and data quality. To bridge this gap, we construct two datasets, namely EduDialogue and SciFormula, following a standardized data construction pipeline with the involvement of educational experts to ensure data quality and procedural consistency [Tirumala *et al.*, 2023].

Formula Collection and Normalization

To ensure broad domain coverage, we collect formulas and multi-turn teacher-student dialogues from an intelligent tutoring system developed by TAL Education Group, a leading educational technology company in China, following standard protocols for data collection and privacy protection. The data covers mathematics, physics and chemistry. All dialogue content is anonymized to remove personally identifiable information while preserving educational relevance. We apply rigorous data cleaning to remove non-scientific, malformed or incomplete formulas. Preprocessing includes whitespace normalization, HTML tag removal, emoji and special character stripping and n-gram-based deduplication to eliminate redundant content. Additionally, all Arabic numerals are converted into textual form to enhance consistency and speech synthesis quality. Valid formulas are then normalized and converted into LaTeX format for downstream processing.

Verbalization and Human Annotation

To obtain accurate speech representations of scientific formulas, we leverage the language understanding capabilities of LLMs to interpret symbolic expressions and generate standardized spoken language. Specifically, we design carefully crafted prompts to instruct Claude Sonnet 4 to transform textual inputs containing formulas into natural and fluent verbalized forms suitable for speech synthesis. For the multi-turn dialogue dataset EduDialogue, every utterance from teachers and students is converted into fluent and complete spoken language to ensure natural continuity in conversation. For the scientific formula dataset SciFormula, each scientific expression in LaTeX format is individually processed to preserve both semantic meaning and structural accuracy. To ensure the quality and correctness of the verbalization, all generated outputs are manually reviewed and refined by educational experts to mitigate potential biases introduced during model generation.

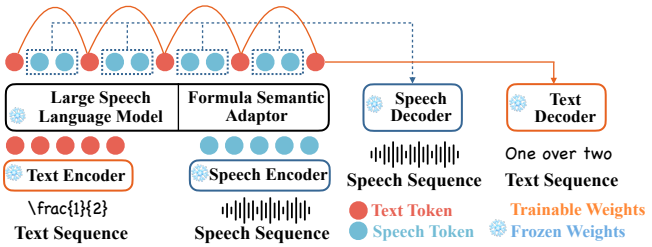

Figure 2: Overview of the proposed Formula-Speech.

Speech Synthesis

To generate high-quality speech data, we synthesize speech from the spoken language representations obtained in the previous step using CosyVoice2 [Du *et al.*, 2024], a state-of-the-art TTS model known for its clarity and expressiveness. The model takes the standardized spoken language as input and produces corresponding speech waveforms. This step enables the construction of paired text-speech data with natural and accurate pronunciation, supporting effective training of LSLMs for scientific formula verbalization. All synthesized speech is manually reviewed by educational experts and any samples judged to be unsatisfactory are regenerated.

In total, we construct two datasets for scientific formula verbalization. The EduDialogue dataset contains 3417 multi-turn dialogues between teachers and students. These dialogues cover subject-specific instructional scenarios and include both speech and the corresponding spoken transcripts. On average, each dialogue has 15.15 turns, with the longest dialogue reaching 40 turns. The distribution of dialogue lengths is diverse, with the highest proportions concentrated at 2 turns accounting for 19.2 percent, 12 turns accounting for 13.7 percent and 14 turns accounting for 13.3 percent. The SciFormula dataset consists of 8678 paired samples spanning mathematics, physics and chemistry. Among them, 4339 samples are in Chinese and 4339 samples are in English. In terms of subject distribution, chemistry contributes 5434 samples which is 62.6 percent of the total, mathematics contributes 2506 samples which is 28.9 percent and physics contributes 738 samples which is 8.5 percent.

4.2 Model Architecture

We build our Formula-Speech upon the GLM-4-Voice backbone to improve scientific formula verbalization in both speech and text modalities [Zeng *et al.*, 2024]. The architecture is designed to align symbolic formula representations with speech while preserving general conversational capabilities. As illustrated in Figure 2, the model takes either speech or text as input and produces outputs in an interleaved text-speech manner across both modalities.

Speech and Text Encoder

Given an input speech sequence $\mathbf{S}$ or a text sequence $\mathbf{T}$, the model first transforms them into discrete representations. A pretrained speech encoder Enc^{sp} maps the speech input into a sequence of speech tokens, denoted as $\mathbf{t}_{sp}$, while another text encoder Enc^{text} processes the text input into a sequence of text tokens $\mathbf{t}_{text}$:

$$\mathbf{t}_{sp} = \text{Enc}^{sp}(\mathbf{S}), \quad \mathbf{t}_{text} = \text{Enc}^{text}(\mathbf{T})$$

Formula Semantic Adaptor

To enable fine-grained multimodal alignment between symbolic structures and their spoken representations, we introduce a lightweight formula semantic adaptor (FSA) that integrates into the attention layers of a frozen LSLM. The FSA is implemented using low-rank parameter updates [Hu *et al.*, 2022], which specialize the query, key, and value projections for symbol-aware speech modeling while keeping all original parameters θ fixed. We denote the resulting model as $\text{LSLM}_\theta^{\text{FSA}}$.

$$(\tilde{\mathbf{t}}_{sp}, \tilde{\mathbf{t}}_{text}) = \text{LSLM}_\theta^{\text{FSA}}(\mathbf{t}_{sp}, \mathbf{t}_{text})$$

where $\tilde{\mathbf{t}}_{sp}$ and $\tilde{\mathbf{t}}_{text}$ represent the aligned output token sequences generated by the LSLM.

Speech and Text Decoder

The aligned token sequences are then passed to two modality-specific decoders. The speech decoder Dec^{sp} generates the output speech $\mathbf{y}_{sp}$ in an autoregressive manner and the text decoder Dec^{text} produces the output text $\mathbf{y}_{text}$:

$$\mathbf{y}_{sp} = \text{Dec}^{sp}(\tilde{\mathbf{t}}_{sp}), \quad \mathbf{y}_{text} = \text{Dec}^{text}(\tilde{\mathbf{t}}_{text})$$

where $\mathbf{y}_{sp}$ denotes the predicted speech sequence for the verbalized formula and $\mathbf{y}_{text}$ denotes the text segment in the same interleaved output that corresponds to this speech sequence.

4.3 Training Strategy

To enable Formula-Speech to learn accurate verbalization of scientific formulas with only a small amount of data, we adopt a two-stage LoRA training framework. This lightweight strategy allows the model to efficiently acquire formula-to-speech alignment patterns while preserving the general language understanding capabilities of the backbone LSLM. This training framework is inspired by recent advances in RL and adaptation methods for LLMs, such as direct preference optimization (DPO) [Ouyang *et al.*, 2022; Rafailov *et al.*, 2023] and GRPO [Shao *et al.*, 2024]. We introduce several modifications tailored to the formula verbalization task, including custom reward functions and a simplified yet effective training pipeline.

Stage I: Supervised Fine-Tuning with LoRA

In the first stage, we apply SFT with LoRA on our proposed datasets to guide the model in learning basic alignment patterns between scientific symbolic inputs and their corresponding spoken outputs.

Let $\mathcal{D}_{\text{SFT}} = (x_i, y_i)_{i=1}^N$ denote the dataset, where x_i is the symbolic formula input and y_i is the corresponding spoken output in the form of speech tokens. We optimize the model using the standard maximum likelihood estimation objective:

$$\mathcal{L}_{\text{SFT}}(\omega) = -\sum_{i=1}^{N} \log P_\omega(y_i \mid x_i)$$

where $\mathcal{L}_{\text{SFT}}$ denotes the SFT loss, ω represents the LoRA adapted model parameters, N is the number of training examples, (x_i, y_i) refers to the i-th symbolic-spoken pair with x_i as the symbolic formula input and y_i as the corresponding speech token output and $P_\omega(y_i \mid x_i)$ is the conditional likelihood of generating y_i given x_i under the model.

Stage II: RL with Custom Rewards

To further enhance output quality and reduce errors such as token duplication or overly lengthy generations, we adopt an RL stage based on policy-gradient optimization. We use the GRPO training framework implemented in TRL[1], which samples multiple candidate outputs, normalizes their reward values within each group and updates the policy using a weighted log-likelihood gradient. This lightweight design improves stability and efficiency for speech generation tasks.

Our RL objective maximizes the expected reward over sampled outputs:

$$\mathcal{L}_{\text{RL}}(\omega) = -\mathbb{E}_{\hat{y} \sim P_\omega(\cdot|x)}\left[r(x, \hat{y})\right]$$

where $\mathcal{L}_{\text{RL}}$ denotes the RL loss and $\hat{y}$ is a sampled speech token sequence drawn from the model distribution $P_\omega(\cdot \mid x)$. The scalar reward $r(x, \hat{y})$ evaluates the quality of $\hat{y}$ conditioned on x and provides learning signals for improving verbalization quality.

To estimate gradients, we apply a standard policy-gradient update of the form:

$$\nabla_\omega \mathcal{L}_{\text{RL}} = -\nabla_\omega \log P_\omega(\hat{y} \mid x) \cdot r(x, \hat{y})$$

where the final update used by GRPO incorporates group-wise reward normalization and clipping for improved training stability.

Although the model generally learns the correct formula-to-speech mappings after SFT, it often produces redundant or overly lengthy speech tokens. To address this, we design a composite reward function $r(x, \hat{y})$ as the weighted average of two custom reward components: (1) a reward for non-repetition and length control that penalizes repeated n-grams in speech token outputs, and (2) a verbalization-quality reward computed via an LLM-as-a-Judge mechanism. The latter transcribes generated speech back to text using ASR and evaluates its correctness relative to the reference interpretation.

To penalize redundant or overly lengthy speech tokens, we define a reward function that computes the n-gram duplication rate in the generated token sequence and applies an exponential penalty when the repetition level exceeds a tolerance threshold.

Let $\hat{y} = \{t_1, t_2, \ldots, t_T\}$ denote the generated sequence of speech, where T is the sequence length and t_i is the i-th speech token. We compute the n-gram duplication rate d to

[1]https://github.com/huggingface/trl.

quantify redundancy in $\hat{y}$. An n-gram is a contiguous subsequence of n tokens from $\hat{y}$, i.e., $g = (t_i, t_{i+1}, \ldots, t_{i+n-1})$ for $1 \leq i \leq T - n + 1$. The duplication rate d is defined as:

$$d = \frac{\sum_{g \in \mathcal{G}} \max(\text{count}(g) - 1, 0)}{T - n + 1 + \epsilon}$$

where $\mathcal{G}$ is the set of all n-grams in $\hat{y}$, $\text{count}(g)$ is the number of times n-gram g appears in $\hat{y}$ and ϵ is a small constant added for numerical stability.

The duplication reward r_{dup} is computed as:

$$r_{\text{dup}} = \begin{cases} 1, & \text{if } d \leq \tau \\ 1 - \exp(\alpha \cdot (d - \tau)), & \text{otherwise} \end{cases}$$

and is clipped to the range $[-1, 1]$ during implementation to ensure numerical stability. d denotes duplication rate, τ is the threshold for duplication rate and α is the penalty scale.

To further enhance verbalization quality, inspired by previous work [Ye *et al.*, 2025; Ma *et al.*, 2025], we adopt an LLM-as-a-Judge framework to assess the correctness of generated outputs. Specifically, the generated speech is first converted into waveform and passed through an ASR module to obtain a predicted transcript $\hat{t}$. The transcript is then compared with a reference transcription t^* using a pretrained LLM to evaluate the fidelity of verbalization with respect to the original formula f.

To improve ASR accuracy, we adopt different ASR models based on language. Paraformer-zh is used for transcribing Chinese utterances [Gao *et al.*, 2022], while Whisper-Large-v3 is used for English utterances [Radford *et al.*, 2023]. In addition, we employ Claude Sonnet 4 as the judging model and design task specific prompts to guide its evaluation of verbalization quality.

Let $s \in [0, s_{\max}]$ be the raw score assigned by the LLM. We normalize this score to a reward value in $[-\mu, \mu]$ as follows: $r_{\text{qual}} = 2\mu / s_{\max} \cdot s - \mu$, where s is the model-assigned raw score, μ is the reward scale and $s_{\max}$ is the maximum score the model can assign.

In total, the final reward is computed as:

$$r(x, \hat{y}) = \lambda \cdot r_{\text{dup}}(x, \hat{y}) + (1 - \lambda) \cdot r_{\text{qual}}(x, \hat{y})$$

where $\lambda \in [0, 1]$ is a balancing coefficient. By optimizing this reward under the GRPO training framework with LoRA updates, the model is guided to generate concise and high-quality verbalization.

5 Experiments

5.1 Experimental Setup

Datasets

To effectively train and evaluate our model, we use two datasets including the multi-turn teacher-student dialogue dataset EduDialogue and the scientific formula verbalization dataset SciFormula. For the EduDialogue dataset, we use the entire dataset for SFT. For the SciFormula dataset, we first divide the data by subject into three subsets including Math2k, Physics1k and Chemistry5k. Each subset is then split into an SFT training set, a RL training set and a test set using a 7:2:1 ratio. We ensure that the proportion of Chinese and English examples is approximately 1:1 within each split. All dataset splits are performed randomly.

Baselines

To evaluate the effectiveness of Formula-Speech, we compare it with state-of-the-art end-to-end LSLMs, including GLM-4-Voice [Zeng *et al.*, 2024], Baichuan-Audio [Li *et al.*, 2025], Step-Audio [Huang *et al.*, 2025], Qwen2.5-Omni [Xu *et al.*, 2025], Kimi-Audio [Ding *et al.*, 2025] and VITA-Audio [Long *et al.*, 2025].

Evaluation Metrics

To comprehensively evaluate the quality of scientific formula verbalization, following prior work [Gao *et al.*, 2022; Cho *et al.*, 2025], we adopt both automatic and human-centered metrics:

- WER (Word Error Rate): It measures the transcription accuracy of the generated speech via ASR. Lower values indicate clearer pronunciation and fewer recognition errors. When insertion errors dominate and many extra words are inserted, the WER value can exceed 1.
- BLEU (Bilingual Evaluation Understudy): It captures the n-gram overlap between ASR-transcribed outputs and reference transcripts, reflecting lexical similarity.
- NMOS (Normalized Mean Opinion Score): A normalized variant of the standard MOS that automatically estimates speech naturalness from acoustic features, with a maximum score of 5. In this paper, we use NMOS estimator from [Dubey *et al.*, 2023].
- LLM_S: It evaluates the correctness and relevance of verbalized formulas using Claude Sonnet 4, which scores each output from 0 to 3 based on the original formula, reference transcript and ASR-transcribed result. Scoring criteria are shown in Table 1. This metric naturally accommodates cases where a single formula admits multiple valid verbalizations.
- Human_S: It measures verbalization accuracy through human annotation using the same rubric as LLM_S, with final scores averaged across three raters. Human evaluation further helps mitigate potential biases introduced by LLM based evaluation.
- Human_F: It assesses speech fluency and naturalness based on predefined criteria in Table 1, with ratings averaged from three raters.

Implementation Details

We adopt a two-stage training strategy for the proposed Formula-Speech. In the first stage, we perform SFT using the full EduDialogue dataset and the SFT training set of the SciFormula dataset, training for 20 epochs with a learning rate of 5e-4 under a cosine schedule. The batch size is 1 per device with gradient accumulation of 4. In the second stage, we apply reinforcement learning on the RL training set of SciFormula using the GRPO training framework implemented in TRL. Training runs for 10 epochs with a learning rate of 1e-6, batch size 8 and accumulation steps 8. We generate 8 responses per prompt using top-p sampling (p = 0.95), temperature 0.6 and up to 512 new tokens. LoRA is applied to the query-key-value modules with rank 64, alpha 32 and dropout 0.05. To guide learning, we use two scalar reward components. The repetition suppression reward penalizes frequent

Evaluation Type	Score	Description
LLM$_S$/Human$_S$	3	Completely accurate and concise; the transcript fully conveys the formula meaning.
	2	Minor errors or deviations (e.g., slight substitutions), but the meaning remains clear and acceptable.
	1	Major errors or unrelated content; key elements are missing or altered.
	0	Severe errors or empty output; formula is misunderstood or unreadable.
Human$_F$	3	Speech is natural and indistinguishable from a real human.
	2	Mostly fluent with minor synthetic artifacts or irregular rhythm.
	1	Noticeably robotic or stiff, but still intelligible.
	0	Severely disfluent or jarring robotic speech.

Table 1: Evaluation criteria for accuracy and fluency scoring.

Model	Size	Math2k				Physics1k				Chemistry5k			
		WER(↓)		BLEU(↑)		WER(↓)		BLEU(↑)		WER(↓)		BLEU(↑)	
		en	zh	en	zh	en	zh	en	zh	en	zh	en	zh
Baichuan-Audio	7B	1.0826	1.1222	0.0501	0.0217	2.9983	1.5152	0.0696	0.0223	2.9356	1.8740	0.1000	0.0245
GLM-4-Voice	9B	1.0903	1.6767	0.0531	0.1794	2.1239	3.6858	0.0237	0.0464	3.0769	4.5244	0.0504	0.0690
Kimi-Audio	7B	1.5760	1.2384	0.0646	0.2510	2.1712	2.1285	0.0337	0.1216	1.8417	2.3513	0.0433	0.1398
Qwen2.5-Omni	7B	1.9217	2.2298	0.0367	0.2307	5.5857	8.5023	0.0243	0.0441	7.3110	11.2037	0.0721	0.0701
Step-Audio	130B	1.0241	0.9846	0.2059	0.4210	1.1209	1.0089	0.2994	0.2440	1.8635	1.2860	0.3289	0.2513
VITA-Audio	7B	1.5950	3.6036	0.0522	0.1157	4.6909	7.4772	0.0287	0.0436	5.3944	9.1039	0.0906	0.0631
Formula-Speech (Ours)	9B	**0.4985**	**0.3421**	**0.4341**	**0.7082**	**0.4777**	**0.4183**	**0.4976**	**0.5090**	**0.6731**	**0.6900**	**0.4894**	**0.5397**

Table 2: Performance comparisons in terms of WER and BLEU.

Model	Size	Math2k				Physics1k				Chemistry5k			
		LLM$_S$(↑)		Human$_S$(↑)		LLM$_S$(↑)		Human$_S$(↑)		LLM$_S$(↑)		Human$_S$(↑)	
		en	zh	en	zh	en	zh	en	zh	en	zh	en	zh
Baichuan-Audio	7B	0.1529	0.0255	0.1900*	0.0240*	0.2766	0.0851	0.5830*	0.0830*	0.3029	0.1294	0.2040*	0.2150*
GLM-4-Voice	9B	0.2357	0.7389	0.2380*	0.6190*	0.2979	0.6383	0.3330*	0.9170*	0.2059	0.5118	0.1720*	0.5700*
Kimi-Audio	7B	0.5414	0.9045	0.2380*	0.8570*	0.4043	1.0851	0.2500*	1.1670*	0.4971	0.8588	0.1290*	0.7850*
Qwen2.5-Omni	7B	0.6815	0.8854	0.2140*	0.9290*	0.8936	0.9149	0.5000*	0.6670*	0.7912	0.9118	0.4190*	0.7310*
Step-Audio	130B	1.0764	1.1019	0.9520*	1.1190*	1.1277	0.9787	1.1110*	1.1110*	**1.2500**	1.1735	1.2810*	1.0830*
VITA-Audio	7B	0.1656	0.4522	0.1670*	0.1670*	0.1489	0.5532	0.3330*	0.1900*	0.2235	0.4471	0.1290*	0.4190*
Formula-Speech (Ours)	9B	**1.2994**	**1.5032**	**1.1900***	**1.7140***	**1.2766**	**1.7447**	**2.3330***	**1.6670***	1.1353	**1.4912**	**1.6340***	**1.7420***

Table 3: Performance comparisons in terms of LLM$_S$ and Human$_S$. Results marked with * have Cohen's Kappa $\kappa > 0.65$.

2-grams using a duplication threshold $\tau = 0.05$, an exponential penalty scale $\alpha = 6$ and a small constant ϵ for numerical stability, with values clipped to $[-1, 1]$. The formula verbalization reward is scored by Claude Sonnet 4 based on ASR transcriptions and linearly mapped to $[-\mu, \mu]$ with $\mu = 1$. The final reward is a weighted sum with coefficient λ. All random seeds are fixed to 42 for reproducibility. All experiments are conducted on a cluster with 8 nodes, each equipped with 8 NVIDIA A100 GPUs.

5.2 Overall Performance

Tables 2 - 3 present the overall performance. The best results are highlighted in bold and the second-best results are underlined. The "zh" denotes Chinese and "en" represents English. Based on the tables, we make the following observations: (1) Our proposed Formula-Speech achieves the best performance across nearly all evaluation metrics, demonstrating its strong capability in scientific formula verbalization. For example, on the Math2k dataset in Chinese, our model attains a WER of 0.3421 and a BLEU score of 0.7082 in Table 2, as well as an LLM$_S$ of 1.5032 and a Human$_S$ of 1.7140 in Table 3, significantly outperforming the second-best model, Step-Audio, which scores 0.9846, 0.4210, 1.1019, and 1.1190, respectively. These results suggest that Formula-Speech is particularly effective at aligning symbolic formulas with their corresponding speech patterns, even in complex domains such as mathematics. (2) Step-Audio also demonstrates competitive performance, ranking second across all metrics and consistently outperforming the remaining baselines. This can largely be attributed to its substantial parameter count of 130B, which is considerably larger than those of other models such as Baichuan-Audio (7B) and GLM-4-Voice (9B). These results suggest that larger model capacity tends to yield better performance in scientific formula verbalization. However, in real-world applications, deploying extremely large models may result in significant resource consumption and higher inference costs, which limits their practicality.

5.3 Comparisons with Cascaded Pipeline

Previous approaches to formula verbalization have predominantly relied on cascaded pipeline where ASR, LLM and TTS modules are connected sequentially. Although effective, these systems often suffer from high latency and cumulative error propagation across stages. Recent research has increasingly focused on end-to-end LSLMs that provide lower latency, enable streaming generation and offer greater potential for real world deployment. To assess the effectiveness of our end-to-end Formula-Speech, we compare it with a strong

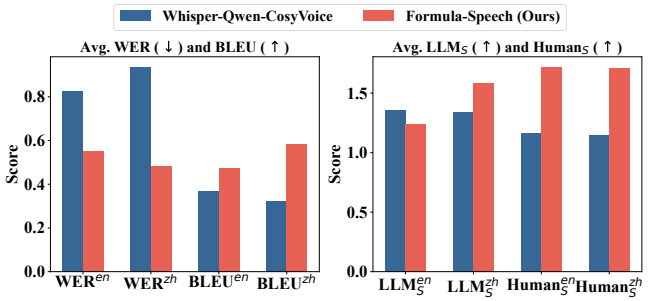

Figure 3: Performance comparisons with cascaded pipeline.

cascaded pipeline using four averaged evaluation metrics including WER, BLEU, LLM$_S$ and Human$_S$. The cascaded pipeline integrates Whisper-Large-v3 for ASR, Qwen2.5-7B-Instruct for text generation and CosyVoice2 for TTS synthesis. Experimental results are shown in Figure 3. The LLM$_S^{en}$ and LLM$_S^{zh}$ denote the metric LLM$_S$ in English and Chinese respectively; others follow accordingly. As illustrated in the figure, Formula-Speech consistently outperforms the cascaded pipeline across all metrics. These findings demonstrate that our end-to-end approach not only enhances verbalization performance but also provides a practical alternative to traditional cascaded pipelines for scientific formula verbalization.

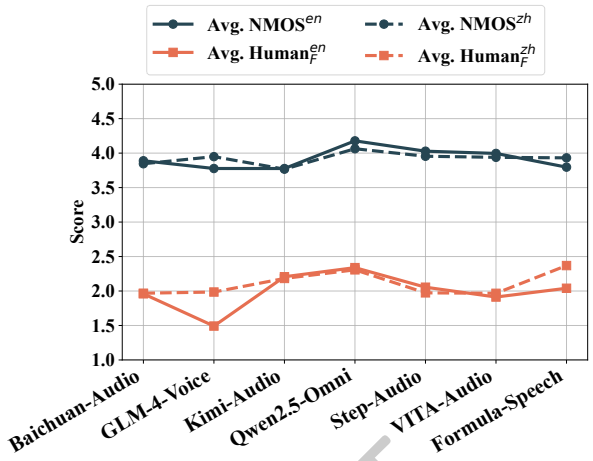

Figure 4: Speech quality comparisons with LSLMs.

Training Strategy	Avg. WER (↓)	Avg. BLEU (↑)	Avg. LLM$_S$ (↑)
Full-SFT	0.5720	0.4976	1.2040
Full-SFT+GRPO	0.5368	0.4544	1.2663
LoRA-GRPO	2.5814	0.0710	0.4625
LoRA-SFT	**0.5039**	0.5203	1.3709
LoRA-SFT+GRPO	0.5166	**0.5297**	**1.4084**

Table 4: Comparisons of different training strategies.

5.4 Speech Quality Evaluation

To assess the overall quality of speech generated by LSLMs, we report the average NMOS scores and collect human ratings on fluency. The evaluation results are presented in Figure 4. From the figure, we observe that most LSLMs exhibit similar speech quality. Among all models, Qwen2.5-Omni achieves the highest average scores on both NMOS and Human$_F$ across Chinese and English tasks. This performance is attributed to the Talker module in its architecture, which enhances the naturalness and smoothness of speech by effectively cloning speaker characteristics. In comparison, Formula-Speech and GLM-4-Voice yield similar results. This outcome is expected since Formula-Speech is built upon GLM-4-Voice without additional optimization for speech fluency or naturalness. Improving speech quality through targeted fluency and voice naturalness enhancement will be explored in future work.

5.5 Impact of Different Training Strategies

To effectively align LSLMs with limited text-speech paired data for scientific formula verbalization, we adopt a two-stage LoRA based training strategy in our Formula-Speech. To verify the effectiveness of this approach, we conduct an ablation study comparing five different training strategies. Full-SFT refers to SFT with all model parameters updated. LoRA-SFT performs SFT using LoRA while keeping the base model frozen. LoRA-GRPO applies RL using the GRPO algorithm with LoRA without prior fine-tuning. Full-SFT+GRPO combines full parameter fine-tuning followed by RL. LoRA-SFT+GRPO represents our complete two-stage training strategy. As shown in Table 4, LoRA-SFT achieves the lowest WER by reducing symbol and structure verbalization errors while preserving the base model's speech generation ability. LoRA-SFT+GRPO further improves BLEU and LLM$_S$, with only a small WER increase caused by slight lexical variations introduced during RL optimization. These results suggest that the two-stage strategy balances transcription-level accuracy with higher-level semantic and pedagogical alignment.

6 Conclusion

In this paper, we focus on scientific formula verbalization for AI tutors, which is essential for accessible learning in online learning systems. Built upon two carefully curated datasets, the proposed Formula-Speech is optimized through a two-stage training strategy combining SFT and RL. Experimental results across multiple scientific domains demonstrate that our model significantly improves the verbalization performance of LSLMs. These findings highlight that end-to-end speech models can be efficiently adapted to handle symbolic content, offering a practical path toward more accessible spoken interaction in educational settings.

7 Stakeholder Involvement

This work was shaped by a real need from online AI tutor development in collaboration with TAL Education Group, a leading educational technology company in China. Existing end-to-end LSLMs often fail to reliably verbalize scientific formulas, limiting their use in AI tutors for accessible learning. Educational experts helped translate this practical need into a research task by selecting educationally meaningful formula-reading cases, refining verbalizations for semantic correctness and pedagogical appropriateness, and shaping the human evaluation criteria for accuracy and fluency.

Acknowledgments

This work was supported in part by Beijing Natural Science Foundation under Grant No. L252010; in part by NSFC under Grant No. 62477025; in part by Key Laboratory of Smart Education of Guangdong Higher Education Institutes, Jinan University (2022LSYS003); in part by the Guangdong University Student Science and Technology Innovation Cultivation Special Fund under Grant No. pdjh2026ak027; in part by Guangdong Hong Kong Joint Laboratory for Data Security and Privacy Protection under Grant No. 2023B1212120007 and in part by the Fundamental Research Funds for the Central Universities.

References

[Abram, 2025] Nicholas Abram. Personalized learning in action: Exploring ai and robotics for early childhood education. In *Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence*, Philadelphia, PA, USA, February 2025.

[Anderer *et al.*, 2025] Katharina Anderer, Karin Muller, Lukas Strobel, Matthias Wolfel, Jan Niehues, and Kathrin Gerling. Making lecture videos accessible for students who are blind or have low vision through ai-assisted navigation and visual question answering. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility*, New York, NY, USA, October 2025.

[Chen *et al.*, 2025] Zui Chen, Tianqiao Liu, Mi Tian, Qing Tong, Weiqi Luo, and Zitao Liu. Advancing mathematical reasoning in language models: The impact of problem-solving data, data synthesis methods, and training stages. In *Proceedings of the 13th International Conference on Learning Representations*, Singapore, April 2025.

[Cho *et al.*, 2025] Cheol Jun Cho, Nicholas Lee, Akshat Gupta, Dhruv Agarwal, Ethan Chen, Alan W Black, and Gopala K Anumanchipalli. Sylber: Syllabic embedding representation of speech from raw audio. In *Proceedings of the 13th International Conference on Learning Representations*, Singapore, April 2025.

[Ding *et al.*, 2025] Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-Audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

[Dong *et al.*, 2022] Su Dong, Shan Liu, Sicen Liu, and Buzhou Tang. Chinese spelling text generation of mathematical formulas. In *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, April 2022.

[Du *et al.*, 2024] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

[Dubey *et al.*, 2023] Harishchandra Dubey, Ashkan Aazami, Vishak Gopal, Babak Naderi, Sebastian Braun, Ross Cutler, Hannes Gamper, Mehrsa Golestaneh, and Robert Aichner. Icassp 2023 deep noise suppression challenge. In *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, Rhodes Island, Greece, June 2023.

[Fang *et al.*, 2025] Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. In *Proceedings of the 13th International Conference on Learning Representations*, Singapore, April 2025.

[Gao *et al.*, 2022] Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Proceedings of the 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, September 2022.

[Gorniak *et al.*, 2024] Joshua Gorniak, Yoon Kim, Donglai Wei, and Nam Wook Kim. Vizability: Enhancing chart accessibility with LLM-based conversational interaction. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, Pittsburgh, PA, USA, October 2024.

[Hadi Mogavi *et al.*, 2023] Reza Hadi Mogavi, Jennifer Hoffman, Chao Deng, Yiwei Du, Ehsan-Ul Haq, and Pan Hui. Envisioning an inclusive metaverse: Student perspectives on accessible and empowering metaverse-enabled learning. In *Proceedings of the 10th ACM Conference on Learning @ Scale*, Copenhagen, Denmark, July 2023.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations*, Virtual Event, April 2022.

[Huang *et al.*, 2025] Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, et al. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*, 2025.

[Hyeon *et al.*, 2025] Sieun Hyeon, Kyudan Jung, Jaehee Won, Nam-Joon Kim, Hyun Gon Ryu, Hyuk-Jae Lee, and Jaeyoung Do. Mathspeech: Leveraging small lms for accurate conversion in mathematical speech-to-formula. In *Proceedings of the 39th Annual AAAI Conference on Artificial Intelligence*, Philadelphia, PA, USA, February 2025.

[Li *et al.*, 2025] Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, et al. Baichuan-Audio: A unified framework for end-to-end speech interaction. *arXiv preprint arXiv:2502.17239*, 2025.

[Liu *et al.*, 2026] Tianqiao Liu, Xueyi Li, Hao Wang, Haoxuan Li, Zhichao Chen, Weiqi Luo, and Zitao Liu. From text to talk: Audio-language model needs non-autoregressive joint training. In *Proceedings of the 14th International Conference on Learning Representations*, Rio de Janeiro, Brazil, April 2026.

[Long *et al.*, 2025] Zuwei Long, Yunhang Shen, Chaoyou Fu, Heting Gao, Lijiang Li, Peixian Chen, Mengdan Zhang, Hang Shao, Jian Li, Jinlong Peng, et al. Vita-audio: Fast interleaved cross-modal token generation for efficient large speech-language model. *arXiv preprint arXiv:2505.03739*, 2025.

[Ma *et al.*, 2025] Yiran Ma, Zui Chen, Tianqiao Liu, Mi Tian, Zhuo Liu, Zitao Liu, and Weiqi Luo. What are step-level reward models rewarding? counterintuitive findings from mcts-boosted mathematical reasoning. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, Philadelphia, PA, USA, February 2025.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Proceedings of the 36th Conference on Neural Information Processing Systems*, New Orleans, LA, USA, December 2022.

[Pal Chowdhury *et al.*, 2024] Sankalan Pal Chowdhury, Vilem Zouhar, and Mrinmaya Sachan. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the 11th ACM Conference on Learning @ Scale*, New York, NY, USA, July 2024.

[Prylipko *et al.*, 2012] Dmytro Prylipko, Björn Schuller, and Andreas Wendemuth. Fine-tuning hmms for nonverbal vocalizations in spontaneous speech: A multicorpus perspective. In *Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 2012.

[Radford *et al.*, 2023] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, HI, USA, July 2023.

[Rafailov *et al.*, 2023] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the 37th Conference on Neural Information Processing Systems*, New Orleans, LA, USA, December 2023.

[Ritchie *et al.*, 2019] Sandy Ritchie, Richard Sproat, Kyle Gorman, Daan van Esch, Christian Schallhart, Nikos Bampounis, Benoît Brard, Jonas Fromseier Mortensen, Millie Holt, and Eoin Mahon. Unified verbalization for speech recognition & synthesis across languages. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, September 2019.

[Roychowdhury *et al.*, 2025] Sujoy Roychowdhury, HG Ranjani, Sumit Soman, Nishtha Paul, Subhadip Bandyopadhyay, and Siddhanth Iyengar. Intelligibility of text-to-speech systems for mathematical expressions. In *Proceedings of the 26th Annual Conference of the International Speech Communication Association*, Rotterdam, Netherlands, August 2025.

[Sak *et al.*, 2013] Hasim Sak, Francoise Beaufays, Kaisuke Nakajima, and Cyril Allauzen. Language model verbalization for automatic speech recognition. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, May 2013.

[Shao *et al.*, 2024] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[Shi *et al.*, 2024] Zhonghao Shi, Amy O'Connell, Zongjian Li, Siqi Liu, Jennifer Ayissi, Guy Hoffman, Mohammad Soleymani, and Maja Mataric. Build your own robot friend: An open-source learning module for accessible and engaging ai education. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, Vancouver, Canada, February 2024.

[Tirumala *et al.*, 2023] Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. D4: Improving LLM pretraining via document de-duplication and diversification. In *Proceedings of the 37th Conference on Neural Information Processing Systems*, New Orleans, LA, USA, December 2023.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, Long Beach, CA, USA, December 2017.

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

[Ye *et al.*, 2025] Ziyi Ye, Xiangsheng Li, Qiuchi Li, Qingyao Ai, Yujia Zhou, Wei Shen, Dong Yan, and Yiqun Liu. Learning llm-as-a-judge for preference alignment. In *Proceedings of the 13th International Conference on Learning Representations*, Singapore, April 2025.

[Zeng *et al.*, 2024] Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. GLM-4-Voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*, 2024.

[Zeng *et al.*, 2025] Aohan Zeng, Zhengxiao Du, Mingdao Liu, Lei Zhang, Shengmin Jiang, Yuxiao Dong, and Jie Tang. Scaling speech-text pre-training with synthetic interleaved data. In *Proceedings of the 13th International Conference on Learning Representations*, Singapore, April 2025.