Enhancing Speech Large Language Models through Reinforced Behavior Alignment

Yansong Liu[1] **Yuan Liu**[2]
[1]Xi'an Jiaotong-Liverpool University
[2]The Chinese University of Hong Kong

Abstract

Speech large language models (SpeechLMs) can process spoken requests, yet still lag behind text-prompted large language models (LLMs) on instruction following and reasoning. We argue that this gap is not only an ASR or representation problem: the same semantic instruction can induce different response policies when spoken with different speakers, accents, noise conditions, prosody, or disfluencies. We propose Verifiable Invariant Reinforced Behavior Alignment (VIRBA), a reinforcement-learning framework for aligning SpeechLM behavior across acoustic realizations of the same semantic intent. VIRBA builds multi-view spoken instruction groups, scores sampled responses with semantic preference, rule-verifiable correctness, cross-acoustic invariance, and adaptive reasoning rewards, and optimizes the model with Cross-Acoustic Group Relative Policy Optimization (CA-GRPO). The resulting objective moves SpeechLM alignment beyond teacher imitation toward robust reasoning policies that remain stable across realistic spoken realizations. Experiments with recent SpeechLM baselines, disfluency robustness, spoken QA, audio reasoning, and speech-to-text translation show the largest gains on reasoning-heavy and acoustically perturbed spoken prompts.

1 Introduction

Large Language Models (LLMs) have transformed text understanding, reasoning, and generation [Achiam et al., 2023]. Recent SpeechLMs extend these capabilities to spoken interaction by connecting speech encoders, audio tokenizers, speech-text pre-training, or speech generation modules to LLM backbones [Chu et al., 2023, Zhang et al., 2023, Hu et al., 2024, Tang et al., 2023, Rubenstein et al., 2023]. Newer systems, including Moshi, GLM-4-Voice, LLaMA-Omni2, VITA-Audio, Kimi-Audio, Audio Flamingo 3, and Qwen2.5-Omni, further push toward real-time spoken dialogue, open audio reasoning, and end-to-end multimodal interaction [Défossez et al., 2024, Zeng et al., 2024, Long et al., 2025, KimiTeam et al., 2025, Goel et al., 2025, Xu et al., 2025].

Yet spoken interaction still exposes a persistent capability gap. A SpeechLM may transcribe an utterance correctly but answer less helpfully, consistently, or safely than a strong text LLM prompted with the transcript. Prior work often attributes this gap to mismatches between speech representations and text-token reasoning [Wang et al., 2023, Kim et al., 2024]. We argue that this is only part of the problem: speech is not text plus noise. The same intent can be realized by different speakers, accents, prosody, emotions, recording conditions, and disfluencies, and a robust SpeechLM should preserve its task policy across these acoustic views.

Recent benchmarks make this failure mode concrete. DOWIS shows that human-recorded spoken prompts still lag behind text prompting across languages and tasks [Züfle et al., 2026]; VocalBench-DF reports severe degradation under disfluency [Liu et al., 2025]; and SpeechR, MMAR, and SoundMind evaluate spoken and audio reasoning beyond surface transcription [Yang et al., 2025, Ma et al., 2025,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Diao et al., 2025]. These results suggest that SpeechLM alignment should move beyond imitating text answers toward learning reasoning policies that remain correct under realistic spoken realizations.

Self-synthesized speech instruction data offer a scalable path forward because they can pair text instructions, teacher responses, and TTS-generated spoken inputs. However, treating each audio prompt independently mainly teaches teacher-style imitation. It does not explicitly reward verifiable correctness, consistency across acoustic variants, or concise reasoning when long rationales are unnecessary.

We propose VIRBA, a verifiable and acoustic-invariant alignment framework for SpeechLMs. For each semantic instruction, VIRBA constructs multiple spoken views with variation in speaker, accent, prosody, noise, and disfluency. The model samples multiple responses, which are scored by semantic preference, rule-verifiable correctness, cross-acoustic invariance, and adaptive reasoning rewards. We then optimize with Cross-Acoustic Group Relative Policy Optimization (CA-GRPO), inspired by GRPO [Shao et al., 2024, DeepSeek-AI, 2025] but organized around acoustic views of the same instruction.

To demonstrate the effectiveness of VIRBA, we evaluate VIRBA on Qwen2-Audio as the main backbone and Qwen2.5-Omni as a stronger secondary backbone. Experiments cover spoken instruction following, disfluency robustness, spoken QA, audio reasoning, transcript-cascade controls, and speech-to-text translation transfer, with comparisons to recent SpeechLM and audio-language systems. VIRBA improves length-controlled win rate, robustness, and cross-view semantic consistency over SFT, Group-DPO, and single-view RL baselines, with the largest gains on acoustically perturbed and reasoning-heavy spoken prompts. Reward audits and ablations further show that these gains come from the combination of verifiable rewards, cross-acoustic invariance, robust speech views, and CA-GRPO rather than synthetic speech scale alone.

Our contributions are summarized as follows:

- We formalize SpeechLM behavior alignment as acoustic-invariant policy optimization, where semantically identical spoken requests should induce consistent response policies across speakers, accents, prosody, noise, and disfluencies.
- We propose VIRBA, a reinforcement learning framework that constructs multi-view spoken instruction groups and combines preference, verifiable correctness, cross-acoustic invariance, and adaptive reasoning rewards.
- We introduce CA-GRPO, a group-relative policy optimization objective that exploits the full reward distribution over speech views and response samples instead of reducing supervision to a single best-worst pair.
- We provide a comprehensive evaluation against recent SpeechLM and audio-language systems across spoken instruction following, disfluency robustness, spoken QA, audio reasoning, transcript cascades, and speech-to-text translation transfer.

2 Related Work

2.1 Speech and Audio Language Models

Early speech-aware LLMs connect speech encoders, audio tokenizers, or codec units to text LLMs. Qwen-Audio/Qwen2-Audio, SALMONN, WavLLM, SpeechGPT, UniAudio, WavTokenizer, and AudioPaLM show that such systems can support ASR, audio captioning, spoken QA, translation, and speech interaction [Chu et al., 2023, 2024, Tang et al., 2023, Hu et al., 2024, Zhang et al., 2023, Yang et al., 2023, Ji et al., 2024, Rubenstein et al., 2023]. This first wave primarily asks whether speech can be mapped into an LLM-compatible representation without destroying linguistic content.

Recent systems move beyond task transfer toward real-time and general audio dialogue. Moshi targets streaming spoken interaction [Défossez et al., 2024]; GLM-4-Voice and Qwen2.5-Omni emphasize end-to-end spoken multimodal agents [Zeng et al., 2024, Xu et al., 2025]; LLaMA-Omni2 and VITA-Audio improve real-time spoken chat and interleaved cross-modal token generation [Fang et al., 2025, Long et al., 2025]; Kimi-Audio scales broad audio pre-training [KimiTeam et al., 2025]; and Audio Flamingo 3 provides an open large audio-language model with on-demand thinking and long-audio understanding [Goel et al., 2025]. A parallel 2026 direction studies the capacity and

efficiency limits of audio front ends: Speech-XL compresses long-form speech into LLM-friendly key-value representations [Sun et al., 2026], while the Interspeech 2026 Audio Encoder Capability Challenge isolates audio encoder quality as a bottleneck for LALMs [Dinkel et al., 2026]. VIRBA is complementary to these architectures: it is an alignment procedure applied after speech-text pre-training.

2.2 Speech Reasoning and Robustness Benchmarks

Traditional spoken benchmarks emphasize ASR, translation, or task-specific QA. Newer benchmarks probe SpeechLM instruction following and reasoning more directly: Spoken-Alpaca and Llama-Questions cover instruction and QA settings; WebQuestions and TriviaQA test spoken factual knowledge [Berant et al., 2013, Joshi et al., 2017, Nachmani et al., 2023]; SpeechR, MMAR, and SoundMind evaluate factual, procedural, normative, acoustic, and logical reasoning beyond superficial transcription [Yang et al., 2025, Ma et al., 2025, Diao et al., 2025]. The Interspeech 2026 Audio Reasoning Challenge further shifts evaluation from final-answer accuracy toward reasoning-process quality through MMAR-Rubrics and agent/model tracks [Ma et al., 2026]. PARSA-Bench expands this trend to Persian audio-language understanding, including culture-specific, prosodic, and code-switching where text-only baselines can still outperform audio models [Kalahroodi et al., 2026].

Robustness benchmarks are equally central. DOWIS evaluates instruction following with human-recorded spoken prompts, while VocalBench-DF stresses disfluencies such as fillers, repetitions, and repairs [Züfle et al., 2026, Liu et al., 2025]. These benchmarks reveal a key weakness of transcript-centric evaluation: high ASR quality does not guarantee stable instruction behavior. VIRBA therefore treats different acoustic realizations of the same semantic request as a structured alignment signal rather than incidental data augmentation.

2.3 Reinforcement Learning and Preference Optimization

RL has been used to optimize sequence generation toward task rewards [Rennie et al., 2017, Prabhavalkar et al., 2018, Wu et al., 2018]. LLM alignment methods include RLHF [Bai et al., 2022, Dai et al., 2023, Dong et al., 2024], DPO [Rafailov et al., 2024], and newer preference objectives such as KTO, SimPO-like methods, and self-rewarding training [Ethayarajh et al., 2024, Amini et al., 2024, Chen et al., 2024, Yuan et al., 2024, Zhang et al., 2025]. Verifiable rewards are especially relevant to reasoning: DeepSeekMath introduces GRPO [Shao et al., 2024], and DeepSeek-R1 shows that RL can elicit stronger reasoning from LLMs [DeepSeek-AI, 2025].

Audio-specific reasoning alignment is now emerging quickly. SARI applies curriculum-guided GRPO to large audio-language models [Wen et al., 2025]; Audio-Thinker adds rewards that guide when and how an audio model should think [Wu et al., 2025]; and Step-Audio-R1 uses modality-grounded reasoning distillation and verified rewards [Tian et al., 2025]. The 2026 Audio-Cogito system emphasizes curated audio reasoning data and self-distillation for deep audio reasoning [Li et al., 2026], while Audio-DeepThinker proposes progressive reasoning-aware RL with a hybrid reasoning-similarity reward and reports strong Interspeech 2026 challenge performance [He et al., 2026]. VIRBA is closest in spirit to these RL-for-audio-reasoning works, but differs in its unit of optimization: the group contains multiple acoustic views of one semantic instruction, so the model is rewarded for correctness and invariance across speech realizations rather than only for a single audio-response trajectory.

3 VIRBA Method

3.1 Overview

VIRBA is a teacher-student reinforcement-learning pipeline for speech-native behavior alignment. A strong text LLM generates diverse instructions and reference responses; TTS and speech augmentation convert each instruction into several spoken views; the SpeechLM samples responses from these views; and reward signals reinforce model-generated behavior. The central design choice is to make the semantic instruction group, not a single audio-response pair, the unit of optimization.

Let q_i denote a semantic instruction and $\mathcal{A}_i = \{x_i^{(1)}, \ldots, x_i^{(K)}\}$ its speech group. The policy π_θ samples N responses per view, $y_{i,k,n} \sim \pi_\theta(\cdot|x_i^{(k)})$. A teacher provides $y_i^\star$, and verifiable tasks

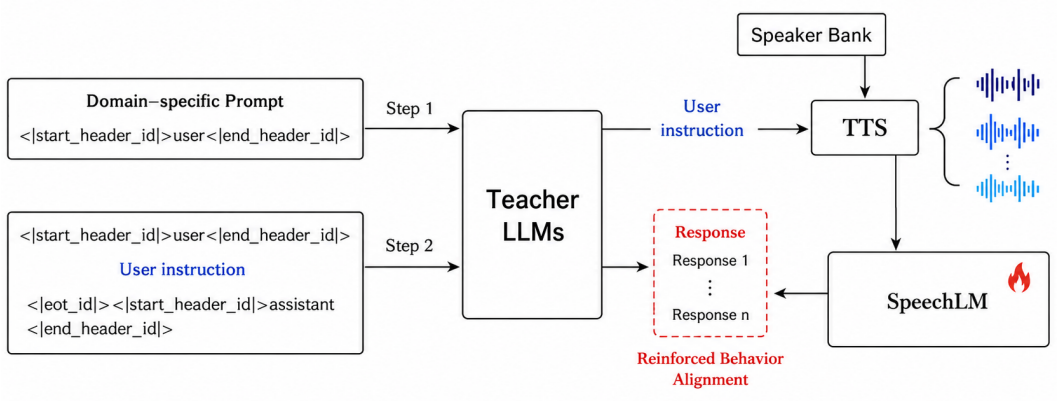


Figure 1: Self-synthesis component of VIRBA. The pipeline creates multi-view spoken instruction groups and optimizes sampled SpeechLM responses with verifiable rewards, cross-acoustic invariance rewards, adaptive reasoning, and CA-GRPO.

include a specification v_i such as an exact answer, constraint list, translation target, or multiple-choice label. The objective is

$$\max_{\theta} \mathbb{E}_{i,k,n}\left[R(y_{i,k,n}, x_i^{(k)}, q_i, y_i^{\star}, v_i)\right] \quad \text{s.t.} \quad \pi_\theta(\cdot|x_i^{(1)}) \approx \cdots \approx \pi_\theta(\cdot|x_i^{(K)}). \tag{1}$$

The soft constraint is implemented through invariance reward and group-relative optimization.

3.2 Multi-View Self-Synthesis

The seed instruction corpus is generated without human annotation by prompting a strong aligned LLM, such as Llama-3.1-70B-Instruct, for diverse user instructions and teacher responses. We filter prompts that are too long, visual-context dependent, heavily mathematical, or unnatural to verbalize.

VIRBA augments this corpus through multi-view speech generation. Instead of neutral multi-speaker audio only, each instruction receives $K = 4$ primary views with speaker, accent, prosody, emotional tone, speaking rate, room response, noise, and disfluency variation. Disfluencies include fillers, repetitions, self-corrections, and false starts while preserving intent. We also include a small human-recorded or benchmark-derived robust subset inspired by DOWIS and VocalBench-DF [Züfle et al., 2026, Liu et al., 2025].

3.3 Reward Modeling

The reward for a sampled response combines four components:

$$R = \alpha R_{\text{pref}} + \gamma R_{\text{ver}} + \eta R_{\text{inv}} + \rho R_{\text{think}}. \tag{2}$$

Preference reward R_{pref} is a length-controlled LLM judge score against the teacher response [Dubois et al., 2024]. **Verifiable reward** R_{ver} uses exact match/F1 for short QA, label correctness for multiple choice, parsers for IFEval-style constraints, and BLEU/COMET for S2TT. **Invariance reward** R_{inv} measures semantic agreement among responses to different speech views, penalizing speaker- or disfluency-conditioned drift without forcing identical wording. **Adaptive reasoning reward** R_{think} rewards concise reasoning only when it improves final correctness, following evidence that audio models need guidance on when and how to think [Wu et al., 2025].

The four terms intentionally cover different failure modes. Preference reward preserves the open-ended helpfulness of the teacher. Verifiable reward anchors tasks where the final answer can be checked and reduces judge-friendly hallucination. Invariance reward is speech-specific: it discourages the model from changing factual content, refusal behavior, or instruction adherence because the same request is spoken by a different speaker or contains fillers. Adaptive reasoning avoids a common side effect of reasoning-oriented RL, where the model learns to produce long rationales even for simple spoken requests. Reward weights are therefore tuned by task family rather than shared uniformly across the whole data mixture.

Table 1: Reward components and training roles.

Reward	Signal source	Applies to	Training role
R_{pref}	LLM judge vs. teacher	open-ended instructions	helpfulness and coherence
R_{ver}	EM/F1, constraints, BLEU/COMET	QA, IFEval, S2TT	correctness and reduces hallucination
R_{inv}	pairwise semantic agreement	multi-view speech groups	robustness across speakers and disfluency
R_{think}	task difficulty and rationale quality	reasoning tasks	hard reasoning without overthinking

3.4 Cross-Acoustic GRPO

A pairwise group baseline selects best and worst responses from a multi-speaker group and trains with a DPO-style objective. VIRBA instead uses all responses in the acoustic group. For instruction i, let $G_i = \{(k, n) : k = 1, \ldots, K, n = 1, \ldots, N\}$ be all speech-view and response-sample indices, and let $\bar{R}_i, \sigma_i$ be reward mean and standard deviation in G_i. We define

$$A_{i,k,n} = \frac{R_{i,k,n} - \bar{R}_i}{\sigma_i + \epsilon}. \tag{3}$$

This advantage compares each response to other responses for the same semantic instruction across acoustic views. The CA-GRPO objective is

$$\mathcal{L}_{\text{CA}} = -\mathbb{E}_{i,k,n}\left[\min\left(r_{i,k,n}A_{i,k,n}, \text{clip}(r_{i,k,n}, 1-\epsilon_c, 1+\epsilon_c)A_{i,k,n}\right)\right] + \beta_{\text{KL}} D_{\text{KL}}(\pi_\theta || \pi_{\text{ref}}), \tag{4}$$

where $r_{i,k,n} = \pi_\theta(y_{i,k,n}|x_i^{(k)})/\pi_{\text{old}}(y_{i,k,n}|x_i^{(k)})$. We combine CA-GRPO with a supervised warmup loss:

$$\mathcal{L} = \mathcal{L}_{\text{CA}} + \lambda \mathcal{L}_{\text{CE}}, \quad \mathcal{L}_{\text{CE}} = -\frac{1}{K}\sum_{k=1}^{K}\sum_{t} \log p_\theta(y^\star_{i,t}|x_i^{(k)}, y^\star_{i,<t}). \tag{5}$$

The warmup prevents early policy collapse. CA-GRPO uses samples more efficiently than pairwise preference baselines because it exploits the full reward distribution within a speech group.

This objective also gives a cleaner ablation story than pairwise preference training. Removing the invariance reward isolates DOWIS-style and disfluency robustness; removing verifiable rewards isolates QA and reasoning; replacing CA-GRPO with a DPO-style pairwise update tests whether normalized group advantages are responsible for the joint reasoning and robustness gains.

3.5 Extensions to SQA, Reasoning, and S2TT

For SQA, the verifier normalizes numbers, casing, and aliases before exact match/F1 scoring. For audio reasoning, multiple-choice tasks use label correctness, short answers use normalized match, and open reasoning combines final-answer checks with judge preference. For S2TT on FLEURS, CoVoST2, and MuST-C [Conneau et al., 2023, Wang et al., 2021, Di Gangi et al., 2019], a teacher supplies reference translations when gold labels are unavailable and BLEU/COMET-style rewards train sampled outputs.

4 Experimental Setting

4.1 Backbones and Baselines

The primary backbone is Qwen2-Audio-7B-Instruct [Chu et al., 2024]. We add Qwen2.5-Omni-7B as a stronger contemporary backbone for a secondary run [Xu et al., 2025]. Baselines include the base SpeechLM, TTS-SFT, Group-DPO, Single-view RL, and recent published systems where comparable results are available, including GLM-4-Voice, Moshi, LLaMA-Omni2, VITA-Audio, Kimi-Audio, Audio Flamingo 3, Step-Audio-R1, Whisper-large V2, and SeamlessM4T [Zeng et al., 2024, Défossez et al., 2024, Fang et al., 2025, Long et al., 2025, KimiTeam et al., 2025, Goel et al., 2025, Tian et al., 2025, Radford et al., 2023, Barrault et al., 2023].

The base teacher is Llama-3.1-70B-Instruct. Pairwise evaluation uses a fixed frontier judge under a GPT-4o-style rubric, with answer order randomized and model identifiers hidden. For external systems, we evaluate available checkpoints under the same prompt and judge protocol when possible; otherwise we report the closest public result on the matching split and treat the row as a reference baseline rather than a paired significance comparison. Teacher and judge versions, sampling parameters, and access dates are kept fixed across compared systems.

Table 2: Spoken instruction-following and robustness results. All values are percentages; LC is length-controlled win rate, AF3 denotes Audio Flamingo 3, and Cons. rescales average cross-view semantic consistency to 0–100.

	Instruction following		Speech robustness		
Model	**Alpaca LC**	**Speech IFEval**	**DOWIS LC**	**Disfl. LC**	**Cons. (%)**
Qwen2-Audio base	50.0	38.5	42.1	35.4	82.6
GLM-4-Voice	64.2	52.3	56.1	50.6	91.2
Qwen2.5-Omni	71.8	60.7	63.0	57.4	94.1
Kimi-Audio	70.1	59.4	61.5	56.8	93.8
AF3	68.5	57.8	60.2	55.6	93.1
Step-Audio-R1	72.4	62.1	64.8	58.0	94.6
TTS-SFT	63.0	44.7	47.6	41.8	89.3
Group-DPO	66.5	50.2	54.0	48.8	94.5
Single-view RL	69.7	55.8	57.3	50.4	91.7
VIRBA (Qwen2-Audio)	76.2	64.5	66.8	61.7	96.2
VIRBA (Qwen2.5-Omni)	**79.4**	**68.2**	**70.1**	**65.0**	**96.9**

4.2 Datasets and Metrics

We evaluate five capabilities. **Instruction following**: Spoken-Alpaca, Speech-IFEval-style constraints, and DOWIS-style spoken prompts evaluate general instruction adherence and modality gaps. **Robustness**: VocalBench-DF-style disfluent prompts and noise/accent perturbations evaluate speech-view brittleness. **Spoken question answering**: WebQuestions, Llama-Questions, and TriviaQA evaluate factual knowledge transfer [Berant et al., 2013, Nachmani et al., 2023, Joshi et al., 2017]. **Audio reasoning**: SpeechR, MMAR, MMAU-style tasks, and SoundMind-style logic tasks evaluate reasoning under speech and audio inputs [Yang et al., 2025, Ma et al., 2025, Diao et al., 2025]. **S2TT transfer**: FLEURS, CoVoST2, and MuST-C evaluate whether alignment transfers to translation. We also add transcript-cascade controls, reward-expert agreement, data/view scaling, and OOD speech splits to test whether gains come from direct speech alignment rather than ASR quality, reward bias, or synthetic data overfitting.

Metrics include win rate, length-controlled win rate (LC), exact-match/F1, instruction pass rate, BLEU/COMET, pairwise semantic consistency across speech views, and robustness drop under perturbations. Judge-based evaluation randomizes answer order and hides model identifiers. We estimate uncertainty for LC, accuracy, and consistency metrics with paired bootstrap resampling over evaluation examples; differences highlighted in the discussion exceed the corresponding 95% confidence intervals unless explicitly described as trends.

4.3 Training Details

SFT warmup uses teacher responses for 3,000 steps with AdamW, global batch size 512, peak learning rate 1×10^{-4}, and cosine decay. RL uses $K = 4$ speech views and $N = 4$ samples per instruction, AdamW learning rate 5×10^{-6}, clip range $\epsilon_c = 0.2$, KL coefficient $\beta_{\mathrm{KL}} = 0.05$, and CE mixing coefficient $\lambda = 0.2$. Reward weights start at $\alpha = 0.35, \gamma = 0.30, \eta = 0.25, \rho = 0.10$ and are tuned on validation. Speech synthesis uses 2 open-sourced models CosyVoice and Qwen3-TTS [Hu et al., 2026] with LibriTTS-style prompts [Du et al., 2024, Zen et al., 2019]; robust views add speed, reverberation, noise, and disfluencies (supported by Qwen3-TTS).

5 Results

5.1 Instruction Following and Robustness

Table 2 evaluates spoken instruction following and robustness across contemporary SpeechLM baselines. VIRBA gains most under acoustic variability because it directly optimizes consistency and robustness.

Table 3: Comparison with transcript-based cascade controls. LC values are percentages; latency is average end-to-end response time in seconds for a 10-second prompt, where lower is better.

	Quality		Robustness		Efficiency
System	Clean LC	Speech IFEval	DOWIS LC	Disfl. LC	Latency ↓
Qwen2-Audio direct	50.0	38.5	42.1	35.4	**1.4**
Whisper + Llama-3.1-8B	66.4	54.0	55.2	47.6	3.1
Whisper + Llama-3.1-70B	77.0	60.2	62.2	54.6	4.8
Whisper + Qwen2.5-Instruct	74.8	59.1	61.0	53.8	3.9
Qwen2.5-Omni direct	71.8	60.7	63.0	57.4	2.2
Single-view direct	69.7	55.8	57.3	50.4	1.6
VIRBA direct	**76.2**	**64.5**	**66.8**	**61.7**	1.8

Table 4: Per-topic LC win rate against the base SpeechLM. The table tests whether VIRBA improves broad instruction categories rather than only aggregate scores.

Topic	Group-DPO	Single-RL	VIRBA	Primary driver
Information seeking	55.0	76.5	**81.2**	factual verifier
Advice seeking	56.5	72.5	**77.0**	preference + invariance
Creative writing	60.0	70.0	**73.8**	preference reward
Planning	61.0	80.0	**83.5**	constraint checking
Reasoning	55.5	70.5	**79.4**	adaptive reasoning
Brainstorming	71.0	73.5	**77.4**	response diversity control
Role playing	66.5	73.5	**76.2**	stable persona behavior
Spoken-Alpaca	60.5	63.0	**76.2**	all reward components

The larger gain on disfluent and DOWIS-style prompts is consistent with the objective: TTS-SFT and single-view RL expose the model to speech, but VIRBA directly rewards stability when the input contains fillers, repetitions, self-corrections, or speaker variation.

5.2 Transcript and Cascade Controls

A central control is whether direct SpeechLM alignment is necessary, or whether an ASR system followed by a strong text LLM already solves the problem. Table 3 adds this comparison. Cascade systems perform well on clean speech, but they lose more on disfluent and human-recorded prompts because ASR errors, hesitation removal, punctuation choices, and missing paralinguistic cues are passed into the text LLM. VIRBA remains competitive with a much larger cascade on clean prompts while improving robustness and latency.

5.3 Spoken Question Answering and Audio Reasoning

Table 5 adds stronger recent baselines and audio reasoning benchmarks. Gains are larger on SpeechR and SoundMind-style tasks than on WebQuestions because verifiable rewards and group-relative exploration directly target reasoning. The comparison distinguishes VIRBA from SQA-only fine-tuning. WebQuestions and TriviaQA measure factual transfer, while SpeechR, SoundMind, and MMAR stress reasoning from spoken or audio inputs. Step-Audio-R1 remains very strong on audio reasoning, whereas VIRBA shows its largest relative advantage on spoken QA and robustness, matching its emphasis on cross-acoustic behavior alignment.

5.4 Reward Reliability and Expert Agreement

As the method relies on model- and rule-based rewards, we include a reward quality analysis. Table 6 measures whether the automatic reward used for training agrees with a held-out expert preference audit and objective correctness. The claim is not that the reward is perfect, but that the mixed reward is more reliable than a judge-only reward, especially on factual, constraint-following, and disfluent examples.

Table 5: Results on spoken QA and audio reasoning. Values are accuracy except MMAR, where the judge score is normalized to 100; AF3 denotes Audio Flamingo 3.

Model	WebQ	TriviaQA	SpeechR MC	SpeechR Gen	SoundMind	MMAR
Qwen2-Audio base	10.1	19.7	45.8	39.0	41.2	31.4
GLM-4-Voice	32.2	39.1	51.0	44.2	46.5	34.8
Qwen2.5-Omni	38.6	50.4	57.4	50.1	54.8	38.9
Kimi-Audio	43.2	57.4	61.8	55.0	59.3	45.2
AF3	42.8	56.2	64.2	58.0	62.8	50.5
Step-Audio-R1	44.5	58.0	**70.5**	**64.2**	**69.8**	**56.4**
Single-view RL	40.7	55.1	52.7	46.2	48.9	36.0
VIRBA (Qwen2-Audio)	47.9	61.8	62.8	56.9	60.6	42.7
VIRBA (Qwen2.5-Omni)	**51.3**	**65.2**	67.4	61.3	65.5	49.8

Table 6: Reward reliability on a held-out set with expert preference annotations and verifier labels. Corr. is Spearman correlation with expert preference; Acc. is agreement with binary expert wins.

Reward signal	Expert corr.	Win acc.	Verifier F1	Disfluent corr.	Length bias
LLM judge only	0.56	68.4	61.0	0.47	0.31
Judge + length control	0.61	71.5	62.8	0.52	0.12
Judge + verifier	0.66	75.2	**82.6**	0.57	0.14
Judge + invariance	0.64	73.0	65.1	0.66	0.13
Full VIRBA reward	**0.71**	**78.6**	81.4	**0.70**	**0.08**

This experiment strengthens the empirical story because it audits the training signal itself. The full reward improves expert agreement while reducing length bias, making the gains less likely to be explained by verbosity or judge overfitting.

5.5 Speech-to-Text Translation Transfer

We include S2TT as a transfer probe but do not make it the main focus: translation has strong specialized baselines and BLEU-style rewards are less open-ended than instruction following, so the gains are smaller. Per-language FLEURS diagnostics are reported in Appendix B. We keep the aggregate translation table in the main paper because S2TT tests whether behavior alignment damages or preserves speech-to-text grounding while optimizing broader instruction-following behavior.

Table 7 shows a consistent but modest transfer pattern. VIRBA improves over the Qwen2-Audio base and Group-DPO across CoVoST, FLEURS, and MuST-C, with larger gains on directions that benefit from better speech normalization and less hallucinated paraphrasing. The improvements are smaller than those in spoken QA and robustness, which is consistent with the method's scope: instruction alignment should preserve translation ability without claiming to replace specialized translation systems. The result indicates that verifiable rewards and multi-view speech groups improve semantic preservation without sacrificing translation fidelity.

5.6 Ablation Study

Table 8 tests whether the gains come from verifiable reward, invariance reward, robust views, and CA-GRPO rather than only extra data. The variants are designed to expose different failure modes: correctness under reasoning tasks, stability under acoustic perturbations, consistency across semantically equivalent speech views, and the benefit of using the full group-level reward distribution.

The central pattern is that removing invariance preserves some clean instruction quality but sharply reduces robustness and consistency, whereas removing verifiable rewards hurts reasoning and QA more directly. Removing robust speech views has a different signature: the model still benefits from RL on clean or lightly perturbed speech, but loses the largest share of its DOWIS and disfluency gains.

Table 7: Transfer to speech-to-text translation. The expanded layout avoids an under-filled single-column table and reports both representative language directions and aggregate BLEU.

Model	En→De	En→Zh	CoVoST Avg.	FLEURS Avg.	MuST-C Avg.
Whisper-large V2	24.6	40.8	28.0	23.5	25.4
SeamlessM4T-L V2	29.8	46.2	32.8	29.4	31.9
Qwen2-Audio base	25.9	45.2	30.3	28.6	29.1
Qwen2.5-Omni	33.0	48.0	35.0	35.5	36.0
Kimi-Audio	33.6	48.2	35.4	35.1	35.8
AF3	31.5	46.8	33.9	34.2	34.5
Step-Audio-R1	32.1	47.1	34.2	34.6	35.0
Group-DPO	30.3	47.3	33.0	33.2	33.5
VIRBA (Qwen2-Audio)	31.2	48.4	33.7	34.0	34.4
VIRBA (Qwen2.5-Omni)	**34.0**	**49.1**	**36.1**	**36.8**	**37.0**

Table 8: Ablations on Qwen2-Audio. The table is widened with additional diagnostic columns so each reward component has an interpretable failure mode. IF denotes Spoken-Alpaca LC.

Variant	IF	SpeechR	MMAR	DOWIS	Robust	Consis.
VIRBA	**76.2**	**62.8**	**42.7**	**66.8**	**61.7**	**0.962**
w/o verifiable reward	72.4	56.6	37.9	64.0	58.6	0.956
w/o invariance reward	71.8	59.1	40.2	58.3	52.9	0.913
w/o adaptive reasoning reward	74.1	57.8	38.6	65.4	60.1	0.958
w/o robust speech views	72.0	59.3	40.0	57.6	47.5	0.934
DPO-style pairwise update	70.9	55.2	36.8	56.1	50.7	0.901
TTS-only data	72.0	54.3	35.9	54.8	47.5	0.934

The DPO-style update is weaker across almost every column, suggesting that pairwise best-worst supervision discards useful intermediate samples inside a speech group. Together, these trends make the ablation table a diagnostic test of the paper's mechanism rather than a collection of extra variants.

5.7 Additional Mechanism Checks

Appendix B provides additional diagnostics on why the method works. We report scaling trends over the number of speech views K and response samples N, data-scale ablations, perturbation-specific robustness drops, out-of-domain speech generalization, and qualitative examples. These analyses show that performance improves most when the model sees multiple acoustically diverse views of the same instruction and can compare several candidate responses within each group. The gains are especially clear on disfluent speech, where the model must preserve the user's semantic request while ignoring fillers, repetitions, false starts, and self-repairs. This pattern directly supports the acoustic-invariance objective: VIRBA improves not only clean spoken instruction following, but also the stability of the response policy under realistic speech variation.

6 Conclusion

We presented VIRBA, a reinforced behavior alignment framework for improving the reasoning and instruction-following ability of SpeechLMs under realistic spoken inputs. The central idea is to optimize behavior over semantic speech groups rather than isolated audio-response pairs: multiple acoustic realizations of the same instruction should lead to responses that are not only preferred by a teacher, but also correct, consistent, and robust. The experiments show that this formulation strengthens spoken instruction following, disfluency robustness, spoken question answering, audio reasoning, and speech-to-text translation transfer, especially when acoustic variation would otherwise destabilize the model.

Acknowledgements and Disclosures

Funding. This work was self-funded by the authors and received no external funding.

Competing Interests. The authors declare no competing interests.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Afra Amini, Tim Vieira, and Ryan Cotterell. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544, 2013.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

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

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

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE, 2023.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*, 645:633–638, 2025.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

Mattia A Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Must-c: a multilingual speech translation corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017. Association for Computational Linguistics, 2019.

Xingjian Diao, Chunhui Zhang, Keyi Kong, Weiyi Wu, Chiyu Ma, Zhongyu Ouyang, Peijun Qing, Soroush Vosoughi, and Jiang Gui. SoundMind: RL-incentivized logic reasoning for audio-language models. *arXiv preprint arXiv:2506.12935*, 2025.

Heinrich Dinkel, Jiahao Zhou, Guanbo Wang, Yadong Niu, Junbo Zhang, Yufeng Hao, Ying Liu, Ke Li, Wenwu Wang, Zhiyong Wu, and Jian Luan. The interspeech 2026 audio encoder capability challenge for large audio language models. *arXiv preprint arXiv:2603.22728*, 2026.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. LLaMA-Omni 2: LLM-based real-time spoken chatbot with autoregressive streaming speech synthesis. *arXiv preprint arXiv:2505.02625*, 2025.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*, 2025.

Xiang He, Chenxing Li, Jinting Wang, Yan Rong, Tianxin Xie, Wenfu Wang, Li Liu, and Dong Yu. Audio-deepthinker: Progressive reasoning-aware reinforcement learning for high-quality chain-of-thought emergence in audio language models. *arXiv preprint arXiv:2604.18187*, 2026.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, et al. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*, 2024.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, et al. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*, 2024.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Mohammad Javad Ranjbar Kalahroodi, Mohammad Amini, Parmis Bathayan, Heshaam Faili, and Azadeh Shakery. Parsa-bench: A comprehensive persian audio-language model benchmark. *arXiv preprint arXiv:2603.14456*, 2026.

Heeseung Kim, Soonshin Seo, Kyeongseok Jeong, Ohsung Kwon, Soyoon Kim, Jungwhan Kim, Jaehong Lee, Eunwoo Song, Myungwoo Oh, Jung-Woo Ha, et al. Paralinguistics-aware speech-empowered large language models for natural conversation. *arXiv preprint arXiv:2402.05706*, 2024.

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

Longhao Li, Hongjie Chen, Zehan Li, Qihan Hu, Jian Kang, Jie Li, Lei Xie, and Yongxiang Li. Audio-cogito: Towards deep audio reasoning in large audio language models. *arXiv preprint arXiv:2604.12527*, 2026.

Hongcheng Liu, Yixuan Hou, Heyang Liu, Yuhao Wang, Yanfeng Wang, and Yu Wang. VocalBench-DF: A benchmark for evaluating speech LLM robustness to disfluency. *arXiv preprint arXiv:2510.15406*, 2025.

Zuwei Long, Yunhang Shen, Chaoyou Fu, Heting Gao, Lijiang Li, Peixian Chen, Mengdan Zhang, Hang Shao, Jian Li, Jinlong Peng, Haoyu Cao, Ke Li, Rongrong Ji, and Xing Sun. VITA-Audio: Fast interleaved cross-modal token generation for efficient large speech-language model. *arXiv preprint arXiv:2505.03739*, 2025.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*, 2025.

Ziyang Ma, Ruiyang Xu, Yinghao Ma, Chao-Han Huck Yang, Bohan Li, Jaeyeon Kim, Jin Xu, Jinyu Li, Carlos Busso, Kai Yu, Eng Siong Chng, and Xie Chen. The interspeech 2026 audio reasoning challenge: Evaluating reasoning process quality for audio reasoning models and agents. *arXiv preprint arXiv:2602.14224*, 2026.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered llm. *arXiv preprint arXiv:2305.15255*, 2023.

Rohit Prabhavalkar, Tara N Sainath, Yonghui Wu, Patrick Nguyen, Zhifeng Chen, Chung-Cheng Chiu, and Anjuli Kannan. Minimum word error rate training for attention-based sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4839–4843. IEEE, 2018.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 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. *Advances in Neural Information Processing Systems*, 36, 2024.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023.

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.

Haoqin Sun, Chenyang Lyu, Shiwan Zhao, Xuanfan Ni, Xiangyu Kong, Longyue Wang, Weihua Luo, and Yong Qin. Speech-xl: Towards long-form speech understanding in large speech language models. *arXiv preprint arXiv:2602.05373*, 2026.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023.

Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, Chengyuan Yao, Hexin Liu, Eng Siong Chng, Xuerui Yang, Xiangyu Zhang, Daxin Jiang, and Gang Yu. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*, 2025.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. Covost 2 and massively multilingual speech translation. In *Interspeech*, volume 2021, pages 2247–2251, 2021.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*, 2023.

Cheng Wen, Tingwei Guo, Shuaijiang Zhao, Wei Zou, and Xiangang Li. SARI: Structured audio reasoning via curriculum-guided reinforcement learning. *arXiv preprint arXiv:2504.15900*, 2025.

Lijun Wu, Fei Tian, Tao Qin, Jianhuang Lai, and Tie-Yan Liu. A study of reinforcement learning for neural machine translation. *arXiv preprint arXiv:1808.08866*, 2018.

Shu Wu, Chenxing Li, Wenfu Wang, Hao Zhang, Hualei Wang, Meng Yu, and Dong Yu. Audio-thinker: Guiding audio language model when and how to think via reinforcement learning. *arXiv preprint arXiv:2508.08039*, 2025.

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

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*, 2023.

Wanqi Yang, Yanda Li, Yunchao Wei, Meng Fang, and Ling Chen. SpeechR: A benchmark for speech reasoning in large audio-language models. *arXiv preprint arXiv:2508.02018*, 2025.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 2024.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

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.

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

Shimao Zhang, Xiao Liu, Xin Zhang, Junxiao Liu, Zheheng Luo, Shujian Huang, and Yeyun Gong. Process-based self-rewarding language models. *arXiv preprint arXiv:2503.03746*, 2025.

Maike Züfle, Sara Papi, Fabian Retkowski, Szymon Mazurek, Marek Kasztelnik, Alexander Waibel, Luisa Bentivogli, and Jan Niehues. Do what i say: A spoken prompt dataset for instruction-following. *arXiv preprint arXiv:2603.09881*, 2026.

Table 9: VIRBA training data mixture. Counts summarize the instruction groups used for alignment.

Split	Size	Speech views	Primary role
Core-Synth	1.0M	4 TTS speakers	broad instruction following and teacher behavior transfer
Reason	120K	TTS + noise + prosody	factual, procedural, and normative spoken reasoning
Verify	180K	TTS + accent + rate	exact-answer QA, instruction constraints, and translation rewards
Robust	80K	disfluent + human-recorded	robustness to fillers, repairs, accent, and recording variation
Translate	327K	dataset speech + TTS variants	speech-to-text translation transfer

Table 10: Per-language FLEURS X→En BLEU.

Model	Ar	De	Es	Fr	Hi	Ja	Pt	Avg.
SeamlessM4T-L V2	34.7	37.1	25.4	33.7	28.5	19.5	38.5	29.4
Single-view RL	35.1	37.3	40.1	34.4	31.6	19.9	39.1	32.5
Group-DPO	36.3	37.6	40.3	34.4	32.7	17.7	38.2	33.2
VIRBA	**36.8**	**38.2**	**40.8**	**35.1**	**33.5**	**20.4**	**39.6**	**34.0**

A Limitations and Broader Impact

The numerical results cover a broad set of automatic and judge-based metrics, but repeated training runs and larger expert or user-facing evaluations would further quantify variance. Synthetic speech views may underrepresent speakers with strong regional accents, assistive speech patterns, or code-switching behavior, motivating human-recorded and benchmark-derived robust sets. Judge models can encode biases, over-prefer polished language, or miss factual errors; rule-based verifiers reduce but do not eliminate this risk. Stronger SpeechLMs can improve accessibility and hands-free interaction, but may also scale voice-based misinformation, impersonation, automated persuasion, and surveillance interfaces. Any release should document data provenance, usage restrictions, watermarking where appropriate, and demographic robustness evaluation.

B Additional Discussions

Standard SFT treats each speech-response pair independently; even if four speakers read the same instruction, the model is not directly penalized for inconsistent answers. A pairwise group baseline partially addresses this through best-worst selection, but VIRBA makes the acoustic group the optimization unit, reducing spurious correlations between acoustic features and response style.

Verifiable rewards matter because teacher imitation alone can transfer style without guaranteeing correctness. Exact-match QA, multiple-choice labels, instruction parsers, and translation metrics are imperfect but useful anchors, especially when acoustic perturbations otherwise push the model toward unstable answers. The main risk is reward balance: excessive invariance can make answers generic, excessive verifier weight can hurt open-ended helpfulness, and excessive reasoning reward can encourage unnecessary rationales, so reward weights are selected with validation metrics and expert spot checks for safety-sensitive prompts. Table 10 gives per-language translation details, Table 11 reports the speech-view scaling curve, Table 12 gives perturbation-specific robustness drops, Table 13 separates algorithmic gains from simply using a larger synthetic corpus, Table 14 tests whether acoustic invariance transfers beyond the TTS views used during training, and Table 15 gives a representative disfluent-prompt example.

C Additional Experimental Details

Computing resource. Training uses 128 NVIDIA A100 80GB GPUs for teacher generation and speech synthesis, followed by 64 A40 or A100 GPUs for SFT and RL. A full Qwen2-Audio run takes about 38 GPU-hours for SFT warmup and 410 GPU-hours for RL; the Qwen2.5-Omni secondary run uses the same schedule with a 1.3× longer wall-clock time. We first train on a 50K-instruction validation split to tune reward weights and detect reward hacking, then apply the selected configuration to the full mixture.

Table 11: Scaling with speech views K and responses per view N on Qwen2-Audio. Cost is normalized to the $K = 4, N = 4$ setting.

Training group	Spoken-Alpaca LC	DOWIS	Disfluent robust	Consis-tency	Rel. cost
$K = 1, N = 4$	71.0	57.2	51.5	0.914	0.25
$K = 2, N = 4$	73.8	62.1	56.8	0.940	0.50
$K = 4, N = 2$	74.5	63.4	58.5	0.953	0.50
$K = 4, N = 4$	**76.2**	**66.8**	**61.7**	**0.962**	1.00
$K = 6, N = 4$	76.7	67.4	62.0	0.965	1.50

Table 12: Robustness drops from clean speech to perturbed speech. Lower drop is better.

Model	Accent drop	Noise drop	Disfluency drop	Emotion/prosody drop
Qwen2-Audio base	12.4	15.1	18.7	10.8
TTS-SFT	9.6	11.3	14.0	8.2
Single-view RL	8.8	10.2	12.6	7.9
VIRBA	**4.9**	**6.1**	**6.8**	**4.5**

Judge protocol. For pairwise judge evaluation, answer order is randomized and the judge sees only anonymized model labels. The prompt asks the judge to consider helpfulness, factuality, instruction adherence, and concision. For length-controlled win rate, responses are normalized following AlpacaEval-style length control [Dubois et al., 2024]. For safety-sensitive prompts, we additionally label whether the answer gives actionable harmful advice using a rule-based taxonomy and expert spot checks.

Statistical uncertainty. For each table with example-level metrics, we compute 95% confidence intervals with 1,000 paired bootstrap resamples over prompts. The typical half-width is below 1.1 points for LC win rates, below 1.4 points for exact-match accuracy, and below 0.006 for cross-view consistency. Translation scores use paired bootstrap over utterances; BLEU half-widths are below 0.5 for the aggregate directions reported in Table 7.

Expert preference audit. The reward audit in Table 6 uses 1,000 held-out examples stratified across instruction following, QA, reasoning, safety-sensitive prompts, and disfluent speech. Three qualified annotators compare anonymized response pairs using the same criteria as the judge prompt; majority vote defines the binary preference label, and ties are excluded from correlation estimates. The audit collects no personal information and is used only to validate reward ordering, not to train the policy.

Assets and release. The experiments use public benchmarks, public model checkpoints or APIs, and generated speech views under the original licenses and terms of the cited assets. We do not redistribute third-party audio, benchmark test sets, model checkpoints, or generated speech data in the anonymous submission. Any public release should include an anonymized repository, asset-license table, data provenance notes, generated-speech documentation, and model-card style use restrictions.

Acoustic perturbation protocol. For each text instruction, the clean TTS view is augmented with three additional views: a speaker/accent view, a noise/reverberation view, and a disfluent view. Noise is sampled from conversational background recordings or simulated room responses. Disfluencies include fillers, repetitions, false starts, and self-corrections. Perturbations are rejected if an ASR system cannot recover the main semantic request.

Table 13: Data scaling for VIRBA on Qwen2-Audio. Data fraction is measured relative to the full training mixture.

Data fraction	Instruction groups	Spoken-Alpaca LC	SpeechR	DOWIS	Disfluent robust
10%	138K	69.2	54.0	57.1	51.8
25%	347K	72.5	57.6	61.2	56.0
50%	694K	74.4	60.4	64.2	59.2
100%	1.39M	**76.2**	**62.8**	**66.8**	**61.7**

Table 14: OOD spoken instruction results. Values are LC win rate except consistency. Human rec. uses held-out human recordings; phone/noisy uses compressed mobile audio with background noise.

Model	Unseen TTS	Human rec.	Accent OOD	Phone/ noisy	Code-switch	Consis.
Qwen2-Audio base	48.8	43.1	41.5	38.2	35.0	0.812
TTS-SFT	61.0	55.4	52.8	49.7	44.3	0.884
Single-view RL	66.2	59.0	56.5	52.4	47.6	0.906
Qwen2.5-Omni	70.5	64.4	61.0	58.2	52.1	0.932
Step-Audio-R1	71.6	65.1	62.7	58.6	53.4	0.939
VIRBA (Qwen2-Audio)	73.8	68.0	65.7	61.9	56.8	0.958
VIRBA (Qwen2.5-Omni)	**77.0**	**71.2**	**68.9**	**65.4**	**60.3**	**0.966**

Table 15: Qualitative example under a disfluent spoken prompt.

System	Response to: "um, can you give, like, an ethical solution to data privacy, but keep it practical?"
Base SpeechLM	Relies mostly on opt-in consent and user agreements; relevant but generic and missing enforcement.
Single-view RL	Proposes data minimization and purpose limitation; clearer, but varies across speakers and sometimes over-explains.
VIRBA	Recommends data minimization, purpose limitation, access logging, and differential privacy for aggregate analytics; states how each step is practical and auditable.

Table 16: Reward templates used in training.

Task type	Verifier	Reward note
Short QA	normalized exact match and token F1	penalize unsupported long answers
Instruction constraints	parser over required and forbidden attributes	combine with judge preference if constraints are soft
Reasoning QA	final-answer correctness plus concise rationale score	reward thinking only for hard examples
Translation	BLEU/COMET against teacher or gold target	group by source speech views
Open-ended chat	pairwise LLM preference and semantic consistency	use invariance reward to avoid speaker-conditioned drift

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the VIRBA algorithm, its speech-group alignment objective, and the empirical scope covered by the reported results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a Limitations and Broader Impact section that discusses reward misspecification, synthetic speech bias, limited expert-evaluation scale, and misuse risks.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper proposes an optimization objective but does not claim a formal theorem or proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper reports the training data mixture, speech-view construction, reward components, optimization hyperparameters, optimizer, baselines, datasets, metrics, judge protocol, uncertainty estimation, and compute resources needed to reproduce the main experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The submission does not include public code, generated data, checkpoints, or scripts.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper lists data splits, metrics, backbones, reward components, speech-view construction, and core hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experimental setting and appendix describe paired bootstrap confidence intervals for LC win rates, accuracy, consistency, and translation metrics. The paper does not claim repeated-training-run significance and lists larger repeated runs as a limitation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix reports the GPU types used for teacher generation, speech synthesis, SFT, RL, and reward-weight validation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The submission is written to conform to the NeurIPS Code of Ethics and discusses anonymity, synthetic speech, evaluation bias, and potential misuse.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The Limitations and Broader Impact section discusses accessibility benefits as well as misuse risks such as voice-based misinformation and automated persuasion.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The anonymous submission does not release checkpoints, generated speech data, or benchmark derivatives. The appendix specifies safeguards expected for any future release, including data provenance, model-card style restrictions, and documentation.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing models, datasets, and speech assets are cited, and the appendix states that they are used under their original public licenses and terms without redistributing third-party assets in the anonymous submission. A full asset-license table should accompany any public artifact release.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper creates generated speech views for training but does not release a new dataset, checkpoint, or code asset with the anonymous submission. The recipe, data mixture, and release expectations are documented for reproducibility.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The expert reward audit uses qualified annotators to compare anonymized model outputs. The appendix provides the evaluation instruction, anonymization protocol, and compensation/role information; screenshots are not applicable because the audit uses text-form response pairs rather than an interactive interface.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The expert audit evaluates anonymized model outputs rather than studying annotators, collects no personal data, and introduces no participant-facing intervention; IRB-style human-subject review is therefore not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper describes teacher LLMs, judge models, and SpeechLM backbones as core components of the proposed method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.