

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 ECHOX: TOWARDS MITIGATING ACOUSTIC-SEMANTIC GAP VIA ECHO TRAINING FOR SPEECH-TO-SPEECH LLMs

Anonymous authors

Paper under double-blind review

ABSTRACT

Speech-to-speech large language models (SLLMs) are attracting increasing attention. Derived from text-based large language models (LLMs), SLLMs often exhibit degradation in knowledge and reasoning capabilities. We hypothesize that this limitation arises because current training paradigms for SLLMs fail to bridge the acoustic-semantic gap in the feature representation space. To address this issue, we propose EchoX, which leverages semantic representations and dynamically generates speech training targets. This approach integrates both acoustic and semantic learning, enabling EchoX to preserve strong reasoning abilities as a speech LLM. Experimental results demonstrate that EchoX, with about six thousand hours of training data, achieves advanced performance on multiple knowledge-based question-answering benchmarks.

1 INTRODUCTION

GPT-4o (Hurst et al., 2024) demonstrates impressive speech interaction performance, which has spurred the rapid development of speech-to-speech large language models (SLLMs). The mainstream approach to building SLLMs is to first discretize speech into speech tokens and then train speech LLMs (Zhang et al., 2023; Défossez et al., 2024; Chen et al., 2025a) under a token-based training paradigm. Although current SLLMs can be trained on massive amounts of speech data, they still exhibit **intelligence degradation** compared to large text-based models (Chen et al., 2024).

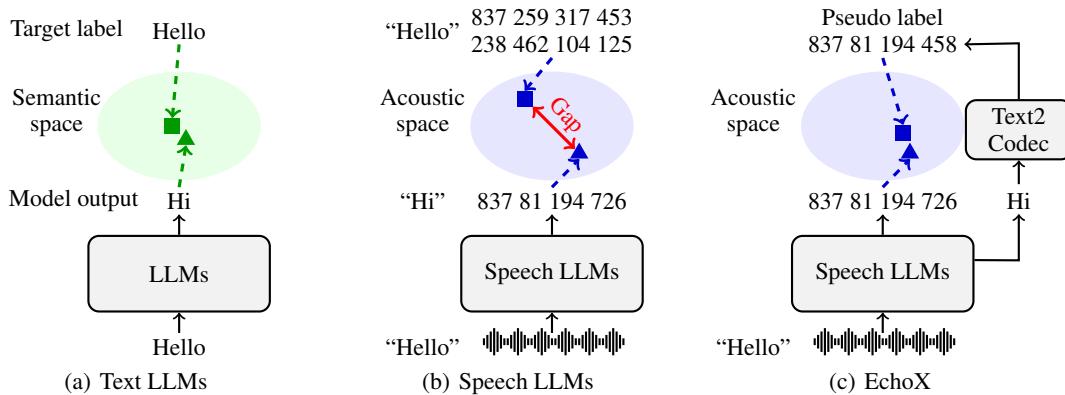


Figure 1: Comparison of training strategies across different models.

Current SLLMs have not yet fully extended the textual intelligence of LLMs into the speech domain, and the underlying reasons for this issue remain underexplored. Beyond the acoustic-semantic conflict in speech

tokens (Gong et al., 2025), we argue that one of the main causes is that SLLMs have not bridged the **acoustic-semantic gap** in the feature representation space. As illustrated in Figure 1(a), the training objective for LLMs emphasizes semantic correctness—predicting a semantically similar token is not heavily penalized. In contrast, SLLMs treat speech tokens as prediction targets, which biases the model toward pronunciation-level accuracy. As a result, even when an SLLM produces a semantically correct response, it may incur severe penalties due to major pronunciation differences, as shown in Figure 1(b).

There are two main paradigms for building SLLMs. The first is interleaved generation (Zeng et al., 2024), which forces the model to jointly consider both acoustics and semantics, but requires a large amount of training data (Chen et al., 2025b). The second employs an auxiliary text-to-codec decoder to convert textual representations into speech tokens (Défossez et al., 2024). However, this approach still fails to address the acoustic-semantic gap.

We propose EchoX, a framework that introduces an auxiliary module to dynamically predict speech tokens based on semantic understanding. This approach eliminates the mismatch between speech tokens and semantic features, enabling the construction of SLLMs that preserve the intelligence of LLMs. Furthermore, to address the challenge of long speech sequences, we adopt *unit language* as the generated speech token and introduce a trigger to support streaming generation, thereby alleviating the difficulties of long-sequence generation. As shown in Figure 2, EchoX achieves advanced performance on knowledge-based QA benchmarks with limited training data and parameters.

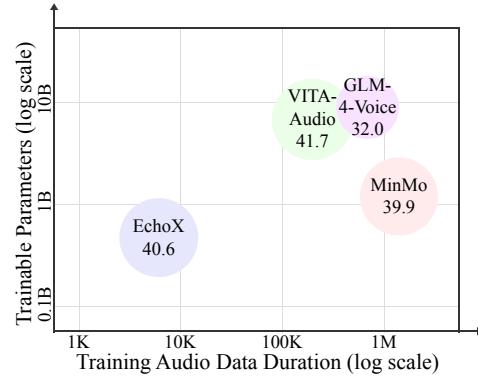


Figure 2: Comparison of models using different training data, parameters, and performance metrics. The number within each node represents the score evaluated on the Web Questions dataset (Berant et al., 2013).

2 METHODOLOGY

2.1 OVERALL DESIGN

We design a three-stage training framework to mitigate the acoustic-semantic gap. The first stage involves converting a textual LLM into a speech-to-text dialog LLM. The second stage trains a text-to-codec model, which converts text into speech tokens. The final stage combines these two modules and fine-tunes the entire speech-to-speech LLM. The overall training process is illustrated in Figure 3. Furthermore, to address the challenge of long speech sequences, we use *unit language* as the speech token and design a streaming inference mechanism.

2.2 STAGE I: SPEECH-TO-TEXT TRAINING

The goal of this stage is to make the model perceive speech and generate textual responses. The mainstream approach involves using an encoder to model the audio, followed by an adapter to bridge the gap between acoustic encoder and textual LLMs (Chu et al., 2024). In our work, we adopt the Soundwave (Zhang et al., 2025a), which employs an alignment adapter and a compression adapter to efficiently achieve audio understanding.

We omit the supervised fine-tuning (SFT) stage described in the original framework, since this work primarily targets spoken dialogue tasks. Instead, we only leverage speech recognition and conversational datasets to build the speech-to-text (S2T) LLM.

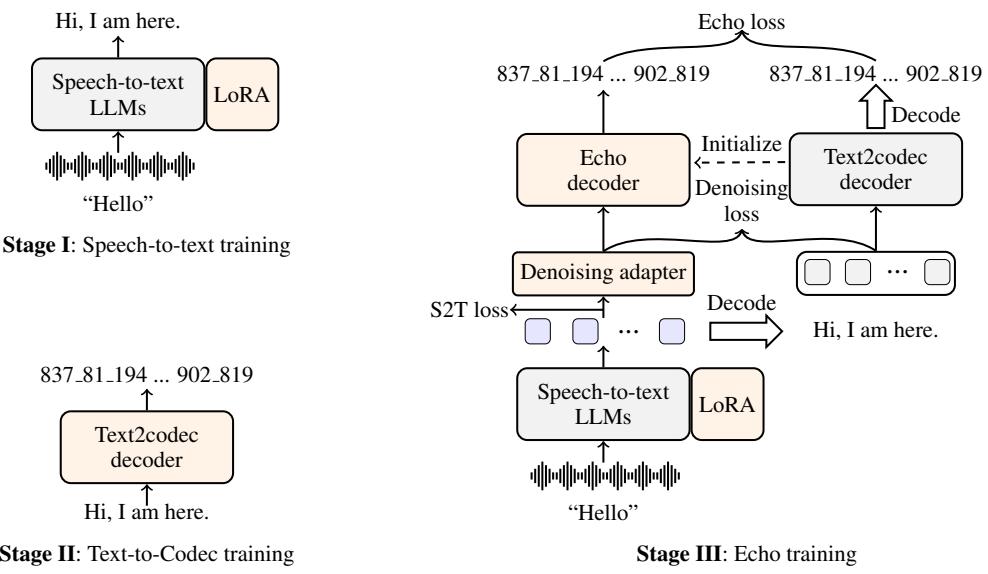


Figure 3: The three training stages of EchoX. The gray modules are frozen while the orange modules are updated. Note that streaming modules are omitted here.

2.3 STAGE II: TEXT-TO-CODEC TRAINING

We use a typical decoder-only architecture to pre-train the text-to-codec (T2C) module (Wang et al., 2023). The input data is text $X = \{x_1, x_2, \dots, x_n\}$, and the target is the sequence of quantized speech tokens $Y = \{y_1, y_2, \dots, y_m\}$. During training, the decoder maps X to hidden states and predicts the speech tokens. We apply a cross-entropy loss for optimization. To ensure consistency of the representation space between the T2C module and the speech-to-text LLM, we initialize and freeze the embeddings, then apply a projection layer to adapt the dimensionality from the LLM to the T2C module.

2.4 STAGE III: ECHO TRAINING

The key objective of this phase is to feed the hidden states from the S2T LLM into the T2C module to generate speech tokens as output. Unlike conventional approaches that rely on annotated speech tokens for training, we propose *Echo training*, which leverages the pre-trained T2C module to decode the outputs of the S2T LLM as training targets.

Formally, let the intermediate representation of the response from the S2T LLM be denoted as $H = \{h_1, \dots, h_n\}$. We perform greedy search to obtain the corresponding text sequence $X' = \{x'_1, \dots, x'_{n'}\}$, which is then fed into the pre-trained T2C module to produce $Y' = \{y'_1, \dots, y'_{m'}\}$ as the final pseudo-labels. During this stage, the T2C module remains frozen.

Echo loss We feed H into an Echo decoder, which shares the same architecture as the T2C module. The Echo decoder is initialized with the T2C parameters. The training objective is to predict Y' , with the loss function defined as:

$$\mathcal{L}_{\text{Echo}} = \sum_i^{m'} \log P(y'_i | H, y'_{<i}) \quad (1)$$

141 Since the hidden states contain redundant information, we design a feed-forward network, termed the De-
 142 noising Adapter, before feeding them into the Echo decoder. The purpose is to align the representations
 143 between H and the embeddings of X' . We employ a cosine similarity loss to train H against X' , thereby
 144 minimizing noise in H and reducing its impact on speech token generation. The training objective is:
 145

$$\mathcal{L}_{\text{Denoising}} = \sum_i^{n'} 1 - \text{Cos}(\text{Adapter}(H_i), \text{Emb}(X'_i)) \quad (2)$$

148 where $\text{Adapter}(\cdot)$ denotes the Denoising Adapter, $\text{Emb}(\cdot)$ represents the word embedding layer in the T2C
 149 module, and $\text{Cos}(\cdot, \cdot)$ computes the cosine similarity between two vectors.
 150

151 **Speech-to-text loss** Additionally, we update the LoRA (Hu et al., 2022) parameters in the first stage for
 152 fine-tuning. We utilize the ground-truth text labels $X = \{x_1, \dots, x_n\}$ for training, with the objective:
 153

$$\mathcal{L}_{\text{S2T}} = \sum_i^n \log P(x_i | H_S, x_{<i}) \quad (3)$$

156 where H_S denotes the hidden state of the input speech S . The final training loss combines all three objectives
 157 through weighted summation:
 158

$$\mathcal{L} = \mathcal{L}_{\text{Echo}} + \lambda * \mathcal{L}_{\text{Denoising}} + \mathcal{L}_{\text{S2T}} \quad (4)$$

159 where λ is a scaling factor, since $\mathcal{L}_{\text{Denoising}}$ differs in nature from the other two losses.
 160

2.5 SPEECH TOKEN CONSTRUCTION

162 We use unit language (Zhang et al., 2025b) as the speech token to reduce the length of the speech sequence.
 163 Unit language significantly compresses the audio sequence while ensuring the quality of text-to-speech
 164 synthesis.
 165

166 **Unit** For speech unit extraction, the raw waveform inputs are first passed through a pre-trained HuBERT
 167 model (Hsu et al., 2021), which transforms them into continuous hidden representations. The selected hidden
 168 layer (the 11th layer in this work) is projected into a k-means codebook space. Each vector is assigned to its
 169 nearest cluster centroid, effectively discretizing the representation into a sequence of unit IDs.
 170

171 **Unit Language** We used unit language, which segments sequences of discrete speech units into word-like
 172 tokens based on statistical language modeling principles (Zhang et al., 2025b). Given a sequence of units
 173 u_1, u_2, \dots, u_n , the goal is to segment and group them into a sequence w_1, w_2, \dots, w_m , where each w_j is
 174 composed of at most K contiguous units.

175 We apply dynamic programming to find the optimal segmentation path $\pi(u_{1:i})$ by maximizing the cumulative
 176 log-probability:
 177

$$k_i^* = \arg \max_k (\underbrace{\log P(\pi(u_{[1:i-k]}))}_{w_{[1:j-1]}^*} + \underbrace{\log P(u_{[i-k+1,i]})}_{w_j}). \quad (5)$$

179 where k_i^* determines the optimal number of units to form w_j and $w_{[1:j-1]}^*$ determines the optimal unit
 180 language before w_j . The unit sequence is segmented recursively based on these optimal values k^* .
 181

182 Normalizing units is important to reduce noise in the unit sequence (Lee et al., 2021). We train an encoder-
 183 decoder model based on the original parallel text-unit data. Then, we perform data distillation on the training
 184 set for regularization purposes. Furthermore, we apply adjacent position deduplication to the units to reduce
 185 the token sequence length.
 186

2.6 STREAMING GENERATION

188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
Table 1: Statistics of data usage at different stages

Task	Data	Size	Duration(H)	Stage
ASR	LibriSpeech (Panayotov et al., 2015)	281,241	960	I
ASR	MLS* (Pratap et al., 2020)	723,636	3,000	I
TTS	AudioQA-1M [†]	178,576	989	II
TTS	SpeechInstruct (Zhang et al., 2023)	31,563	84	II
TTS	HH-RLHF-Speech [‡]	124,945	656	II
SQA	sharechatx (Cheng et al., 2025)	43,223	178	I, III
SQA	Magpie-Pro-Speech+ [‡]	117,000	327	I, III
Total	-	1,500,184	6,194	-

[†] AudioQA-1M: text-only usage with minor cleanup; all audio is synthesized by ourselves. Sourced from VITA-1.5 (Fu et al., 2025).

[‡] Speech versions of two public *text-only* conversational datasets—hh-rlhf (Bai et al., 2022) and Magpie-Llama-3.3-Pro-1M-v0.1 (Xu et al., 2024)—created via light text normalization and TTS; For Magpie, we additionally extend the corpus to improve coverage. The speech version of the two datasets are denoted HH-RLHF-SPEECH and MAGPIE-PRO-SPEECH+.

* denotes we sample the dataset and only used part of it.

Note there is no target audio at stage III; thus the duration count only contains source speech.

Given that speech sequences are significantly longer than their text counterparts, waiting for complete text generation before producing speech tokens would substantially increase synthesis difficulty. Therefore, applying streaming generation becomes essential, as it mitigates long-sequence generation challenges and improves real-time responsiveness.

The core of streaming generation lies in determining whether to read (continue processing) or write (start generating speech) at each timestep. The critical constraint is maintaining the semantic completeness of each segment to avoid disjointed speech output.

We implement a trigger feature that computes the cosine similarity between the current semantic representation and The trigger feature. A write operation is executed (sending the subsequence to the Echo decoder) only when similarity exceeds a threshold and the current value is a local extremum of the window size w . The streaming inference process is shown in Figure 4.

3 DATA

To construct high-quality corpora for *Speech-to-Text* (S2T), *Text-to-Codec* (T2C), and *Speech-to-Speech* (S2S) training, we adopt a data-centric pipeline with four stages: (i) collect text dialog corpora suited for spoken interaction; (ii) transform them into natural, spoken-style dialogues via a rigorous multi-step cleaning and rewriting process; (iii) synthesize the required acoustic modalities (inputs and/or outputs) with carefully controlled voices; and (iv) enforce strict audio quality control to retain only reliable samples. Appendix A shows the detailed process for the pipeline. Statistics of the training data are shown in Table 1.

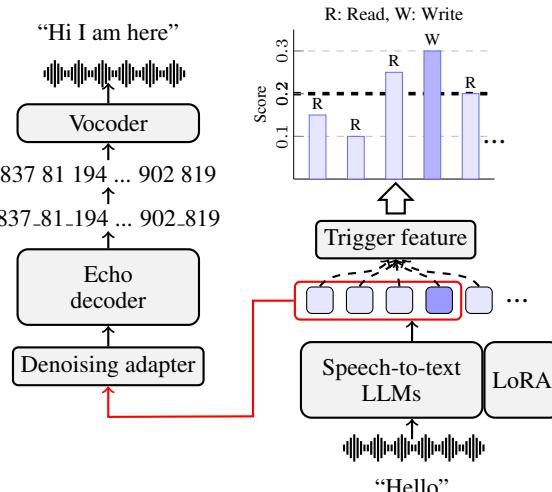


Figure 4: Streaming inference process.

235 3.1 SPEECH-TO-TEXT TRAINING
236237 We apply the above pipeline to a collection of open-source dialog datasets (e.g., *Magpie*), clean them into
238 spoken-style text, and synthesize user-side inputs with diverse Google TTS voices¹. Text-based dialog
239 data typically generates structured and formal outputs, which introduce excessive non-speech tokens (e.g.,
240 symbols, formatting cues). For instance, the token “1.” can be interpreted differently depending on the
241 context—“one point” in mathematical text or “first” in a list.242 To verify acoustic integrity and textual alignment, we transcribe the synthesized inputs using the
243 `parakeet-tdt-0.6b-v2` ASR model and compute word error rate (WER). We retain only utterances
244 with $WER < 5\%$.245 3.2 TEXT-TO-CODEC TRAINING
246247 Using the same cleaning pipeline, we process additional sources including *AudioQA*, *SpeechInstruct*, and
248 *hh-rlhf*. For each assistant turn, we synthesize single-voice speech with the fine-tuned GPT-SoVITS² model
249 and extract codec tokens. The T2C supervision used in training is $\langle \text{text, codec} \rangle$ only, explicitly aligning
250 textual content with its corresponding codec representation.251 To broaden S2S coverage and promote generalization, we also synthesize input speech for the *hh-rlhf* user
252 prompts using the Google TTS API, thereby yielding paired user speech and assistant speech for those
253 dialogs. The resulting S2S dialog sets will be released alongside our corpus.254 3.3 ECHO TRAINING
255256 This portion of data primarily consists of three parts. The first part is everyday dialogue, where the model
257 acts as an assistant, and the overall distribution is relatively short. The second part is speech reasoning,
258 where the input is a speech-based question and the output is a long-text reasoning process. The third part is
259 knowledge-based Q&A data, mainly comprising question-and-answer interactions about common sense.260 4 EXPERIMENTS
261262 4.1 MODEL SETTINGS
263264 We conducted experiments based on two model sizes: 3B and 8B. For the 3B model (called EchoX-3B), we
265 used LLaMA 3.2, while the 8B model (called EchoX-8B) used LLaMA 3.1 (Grattafiori et al., 2024). For
266 the Text2Codec model, both the Echo Decoder and Text2Codec adopted the same architecture: For the 3B
267 model, 6 Transformer layers with a hidden dimension of 512. For the 8B model, 8 Transformer layers with
268 a hidden dimension of 768. For Speech2Speech, an additional adapter was used with an intermediate layer
269 size of 8192. The value of λ to balance the training loss is set to 0.2. The vocoder we used is the unit-based
270 HiFi-GAN (Kong et al., 2020; Polyak et al., 2021).271 For training steps, Stage I: Trained for 10,000 steps, primarily referencing SoundWave (Zhang et al., 2025a).
272 Stage II: Trained for 5,000 steps using 4 GPUs. Stage III: Trained for 12,000 steps—using one 8 A100
273 GPUs for the 3B model and 16 A100 GPUs for the 8B model. We take the embedding of *period* as the
274 trigger representation. The streaming threshold is set to 0.1 and the w for streaming window is set 5. For
275 all our models we use the greedy search to inference. For evaluation, we use the UltraEval-Audio toolkit³.276
277
278
279 ¹<https://cloud.google.com/text-to-speech/docs/list-voices-and-types>²<https://github.com/RVC-Boss/GPT-SoVITS>³<https://github.com/OpenBMB/UltraEval-Audio>

282 We mainly conduct experiments on the three benchmarks: Llama questions (Nachmani et al., 2023), Web
 283 questions (Berant et al., 2013), and TriviaQA (Joshi et al., 2017).
 284

285 Table 2: Speech-to-Speech performance on spoken QA benchmarks.
 286

Model	Llama Questions	Web Questions	TriviaQA	Avg.
OmniDRCA(2B) (Tan et al., 2025)	55.3	22.1	17.9	31.8
LLaMA-Omni2-3B (Fang et al., 2025)	55.7	28.0	-	-
EchoX-3B	54.0	31.6	25.8	37.1
GPT-4o-Realtime (Hurst et al., 2024)	71.7	51.6	69.7	64.4
VITA-Audio (Long et al., 2025)	68.0	41.7	41.7	50.5
MinMo (Chen et al., 2025b)	64.1	39.9	37.5	47.2
MiniCPM-o 2.6 (Yao et al., 2024)	61.0	40.0	40.2	47.1
OmniDRCA (8B) (Tan et al., 2025)	65.0	30.0	32.9	42.6
GLM-4-Voice (Zeng et al., 2024)	50.0	32.0	36.4	39.5
LLaMA-Omni2-7B (Fang et al., 2025)	60.7	31.3	-	-
Freeze-Omni [*] (Wang et al., 2024)	46.0	26.1	25.7	32.6
Moshi (Défossez et al., 2024)	43.7	23.8	16.7	28.1
EchoX-8B	63.3	40.6	35.0	46.3

299 ^{*} indicates that we retested the model using the same evaluation tool.
 300

301 Table 3: Speech-to-Text performance on spoken QA benchmarks.
 302

Model	Llama Questions	Web Questions	TriviaQA	Avg.
LLaMA-Omni2-3B (Fang et al., 2025)	64.3	30.5	-	-
EchoX-3B	73.0	40.8	36.1	50.0
MinMo (Chen et al., 2025b)	78.9	55.0	48.3	60.7
OmniDRCA (Tan et al., 2025)	79.7	51.7	47.7	59.7
VITA-Audio (Long et al., 2025)	75.6	45.0	45.9	55.5
LLaMA-Omni2-7B (Fang et al., 2025)	64.3	30.5	-	-
EchoX-8B	77.3	44.6	46.7	56.2

311 4.2 RESULTS

313 We compared our model with others on knowledge-based question-answering tasks in Tables 2 and 3. It
 314 can be observed that models using the interleave approach, despite being trained on massive amounts of
 315 data, show no significant advantage in speech-to-text tasks—indicating that the core challenge lies in jointly
 316 modeling speech and text representations.

317 For speech-to-speech tasks, although interleave-based models currently demonstrate certain advantages,
 318 models using the T2C method can still efficiently achieve comparable performance. Our proposed EchoX
 319 trained with about six thousand hours of data, achieves comparable performance with models trained on
 320 millions of hours. Thus, our proposed Echo training strategy offers an efficient way to learn unified speech
 321 and semantic representations.

323 5 ANALYSIS

325 We begin by comparing the knowledge degradation in SLLMs and further apply case studies to interpret its
 326 causes from a representational perspective. We then conduct comparative experiments on two approaches
 327 for long-sequence generation: unit language modeling and streaming decoding. We use EchoX-3B for the
 328 analysis unless otherwise specified.

329 5.1 INTELLIGENCE DEGRADATION OF SLLMs
330331 We analyze how knowledge degradation occurs in SLLMs. From the results in Table 4, it can be observed
332 that the Speech-to-Text model improves performance on simple question-answering tasks like LLaMA Ques-
333 tions, but leads to a significant decline on more challenging tasks.334 As for the speech output, even incorporating a TTS model for the S2T model leads to a further decrease,
335 due to errors in synthesizing and recognizing certain specialized nouns. Furthermore, when building an end-
336 to-end model, if an interleaved training strategy is directly adopted, severe knowledge degradation emerges
337 at this data scale. Employing an additional decoder can alleviate this issue by reducing the inconsistency
338 between acoustic and semantic learning, though noticeable interference still persists. By using the Echo
339 decoder, conflicts can be further mitigated, enabling simultaneous learning of both speech and text.
340341 Table 4: Performance comparison of models using the same data and different training strategies.
342343
344
345
346
347
348
349
350

Model	Llama Questions	Web Questions	TriviaQA	Avg.
Text output				
Text-to-text	67.3	53.1	50.0	56.8
Speech-to-text	73.0	40.8	36.1	50.0
Speech output				
Cascade	61.3	37.1	31.3	43.2
Interleaving	21.3	10.6	6.4	12.8
EchoX w/o Echo training	40.3	20.0	12.6	24.3
EchoX	54.0	31.6	25.8	37.1

352 5.2 ACOUSTIC-SEMANTIC GAP
353354 We compare the similarity of word representations across different
355 models in Figure 5. “Hi” and “Hello” are semantically close, while
356 “Hi” and “High” are acoustically similar. It can be observed that
357 in the S2T model, the similarity between “Hi” and “Hello” is rel-
358 atively high. However, after training, the similarity between them
359 decreases. Additionally, the similarity of their speech tokens is very
360 low, essentially indicating no correlation. For “Hi” and “High”,
361 regardless of whether in the S2T model or after interleaving training,
362 their similarity remains relatively low. However, their speech tokens
363 are highly consistent. This demonstrates that the learning objectives
364 for semantics and acoustics are not aligned, necessitating the design
365 of solutions to address this issue.
366367 Table 5: Length ratio and performance comparison of two types of codec. *Length R.* refers to the ratio of
368 speech token to text token.
369370
371
372
373
374
375

Speech token	Llama Questions		Web Questions		TriviaQA	
	Length R. \downarrow	ACC \uparrow	Length R. \downarrow	ACC \uparrow	Length R. \downarrow	ACC \uparrow
Unit	9.31	49.0%	9.63	28.8%	9.13	24.7%
Unit language	4.57	54.0%	4.79	31.6%	4.57	25.8%

5.3 EFFECT OF SPEECH TOKENS

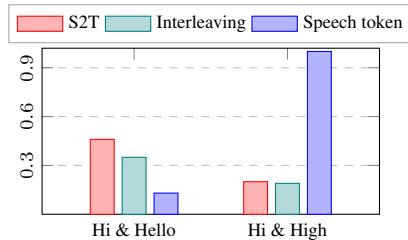


Figure 5: Similarity between two words within different model.

We compared the results of using *unit* and *unit language* as speech tokens in Table 5. It can be observed that using unit language achieves nearly twice the compression ratio while delivering superior performance. Additionally, we further compared the quality of the generated audio and found that both methods perform similarly in terms of audio quality, as shown in Figure 6. However, the recognition accuracy of audio generated with unit language is significantly higher than that generated with units. This also indirectly indicates that when the model predicts speech tokens based on hidden representations, it is prone to error accumulation, leading to an increased error rate in the final model predictions.

387 5.4 EFFECT OF STREAMING INFERENCE

We compared streaming and offline methods under both 3B and 8B model sizes in Table 6. The results show that using a streaming approach does not introduce significant performance degradation. Moreover, at the 3B level, due to the limited capacity of the LLM, properly segmenting the sequences helps the synthesis model achieve better performance and improved results. This demonstrates that streaming decoding reduces the difficulty of generating long sequences.

400 6 RELATED WORK

402 Currently, two mainstream approaches are widely adopted to training SLLMs: interleaving decoding and 403 text2codec decoding.

404 The interleaving method aims to enable the model to learn both audio tokens and text tokens simultaneously, 405 thereby unifying semantic and acoustic representations (Zeng et al., 2024). Although this approach allows 406 direct joint input of speech and text tokens, it requires massive amounts of text and speech data to achieve 407 satisfactory performance (Long et al., 2025; Tan et al., 2025; Li et al., 2025).

408 The alternative method employs an additional text2codec decoder to convert text representations into speech 409 representations (Défossez et al., 2024; Huang et al., 2025; Ding et al., 2025; Chen et al., 2025b). This strategy 410 effectively decouples speech learning from semantic learning, helping to better preserve knowledge while 411 reducing the demand for extremely large training datasets (Fang et al., 2025; Wang et al., 2024). However, 412 few works investigate the causes of intelligence degradation. In this work, we propose Echo Training to 413 bridge the acoustic-semantic gap, enabling more flexible and effective model training.

415 7 CONCLUSION

418 We propose EchoX, which primarily addresses the issue of intelligence degradation in current SLLMs. We 419 first identified that existing training paradigms tend to cause an acoustic-semantic gap. To mitigate this, 420 we introduced the Echo decoder architecture and a corresponding training strategy, and further adopted a 421 more efficient and compact unit language as speech tokens. Experiments demonstrate that our model, using 422 around six thousand hours of data, achieves comparable performance to the model based on millions of hours of data on intelligence QA tasks.

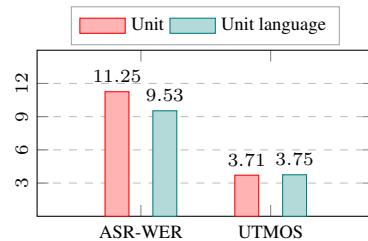


Figure 6: Comparison the speech quality based on two speech tokens.

	Latency (tokens)	Llama Questions	Web Questions	TriviaQA
EchoX-3B				
Streaming	27.17	54.0	31.6	25.8
Offline	138.46	55.3	31.0	24.9
EchoX-8B				
Streaming	29.79	62.0	38.2	31.7
Offline	175.34	64.0	38.3	32.1

Table 6: Performance comparison between streaming and offline decoding methods.

423 REPRODUCIBILITY STATEMENT
424425 We provide a detailed description of the data construction process in Appendix A. Appendices B and C
426 outline the model architecture and training parameters, respectively. Upon peer-review, we will open-source
427 our training data, model, and code to ensure reproducibility.
428429 REFERENCES
430431 Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, et al. Training a helpful and harmless
432 assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
433434 Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-
435 answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*,
436 pp. 1533–1544, 2013.437 Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng,
438 Kuo Yang, et al. Emova: Empowering language models to see, hear and speak with vivid emotions. In
439 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5455–5466, 2025a.
440441 Qian Chen, Yafeng Chen, Yanni Chen, Mengzhe Chen, Yingda Chen, Chong Deng, Zhihao Du, Ruize
442 Gao, Changfeng Gao, Zhifu Gao, et al. Minmo: A multimodal large language model for seamless voice
443 interaction. *arXiv preprint arXiv:2501.06282*, 2025b.
444445 Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. Voicebench: Bench-
446 marking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*, 2024.
447448 Xize Cheng, Dongjie Fu, Xiaoda Yang, Minghui Fang, Ruofan Hu, Jingyu Lu, Jionghao Bai, Zehan Wang,
449 Shengpeng Ji, Rongjie Huang, Linjun Li, Yu Chen, Tao Jin, and Zhou Zhao. Omnichat: Enhancing spoken
450 dialogue systems with scalable synthetic data for diverse scenarios. *arXiv preprint arXiv:2501.01384*,
451 2025. Introduces the ShareChatX dataset.
452453 Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng
454 He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.
455456 Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard
457 Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint
arXiv:2410.00037*, 2024.458 Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan,
459 Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.
460461 Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. Llama-omni2: Llm-based real-time
462 spoken chatbot with autoregressive streaming speech synthesis. *arXiv preprint arXiv:2505.02625*, 2025.
463464 Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long,
465 Heting Gao, Ke Li, Long Ma, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. Vita-1.5:
466 Towards GPT-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025.
467468 Yitian Gong, Luozhijie Jin, Ruiyan Deng, Dong Zhang, Xin Zhang, Qinyuan Cheng, Zhaoye Fei, Shimin Li,
469 and Xipeng Qiu. Xy-tokenizer: Mitigating the semantic-acoustic conflict in low-bitrate speech codecs.
arXiv preprint arXiv:2506.23325, 2025.

470 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-
 471 Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models.
 472 *arXiv preprint arXiv:2407.21783*, 2024.

473

474 Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Ab-
 475 delrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of
 476 hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

477 Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu
 478 Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

479

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

483 Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow,
 484 Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*,
 485 2024.

486

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

489

490 Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient
 491 and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033,
 492 2020.

493

494 Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang,
 495 Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, et al. Textless speech-to-speech translation on real data.
 496 *arXiv preprint arXiv:2112.08352*, 2021.

497

498 Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and
 499 Tatsunori B Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models, 2023.

500

501 Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng
 502 Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.

503

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

507

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

511

512 Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based
 513 on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal
 514 processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

515

516 Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman
 517 Mohamed, and Emmanuel Dupoux. Speech resynthesis from discrete disentangled self-supervised repre-
 518 sentations. *arXiv preprint arXiv:2104.00355*, 2021.

519

520 Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale
 521 multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*, 2020.

517 Chao-Hong Tan, Qian Chen, Wen Wang, Chong Deng, Qinglin Zhang, Luyao Cheng, Hai Yu, Xin Zhang,
 518 Xiang Lv, Tianyu Zhao, et al. Omnidrc: Parallel speech-text foundation model via dual-resolution speech
 519 representations and contrastive alignment. *arXiv preprint arXiv:2506.09349*, 2025.

520

521 Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu,
 522 Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers.
 523 *arXiv preprint arXiv:2301.02111*, 2023.

524

525 Xiong Wang, Yangze Li, Chaoyou Fu, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. Freeze-
 526 omni: A smart and low latency speech-to-speech dialogue model with frozen llm. *arXiv preprint
 527 arXiv:2411.00774*, 2024.

528

529 Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen
 530 Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *arXiv
 531 preprint arXiv:2406.08464*, 2024.

532

533 Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin
 534 Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*,
 535 2024.

536

537 Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and
 538 Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint
 539 arXiv:2412.02612*, 2024.

540

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

544

545 Yuhao Zhang, Zhiheng Liu, Fan Bu, Ruiyu Zhang, Benyou Wang, and Haizhou Li. Soundwave: Less is
 546 more for speech-text alignment in llms. *arXiv preprint arXiv:2502.12900*, 2025a.

547

548

549

550

551

552

553

554

555

556

557

558

559

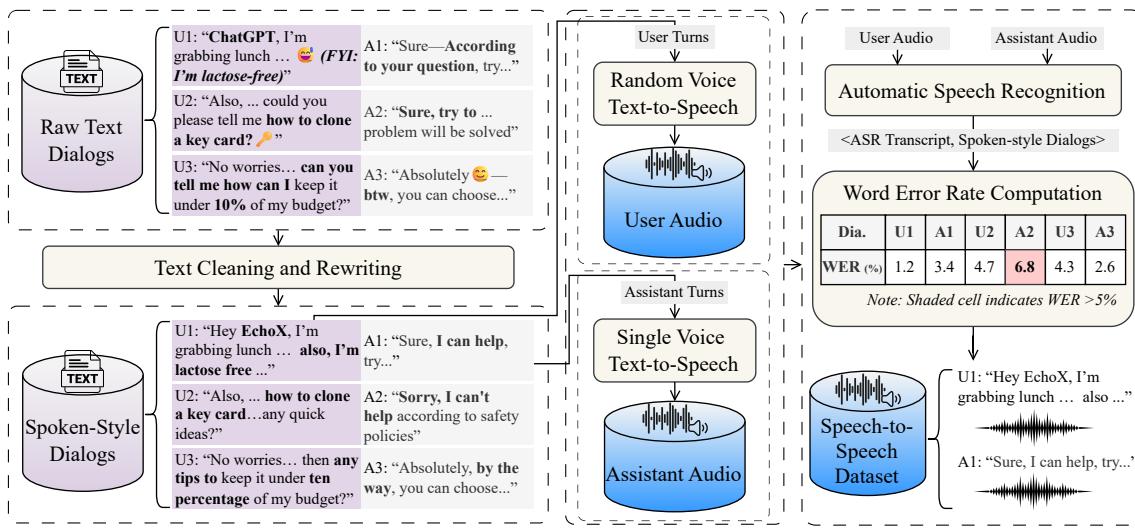
560

561

562

563

Yuhao Zhang, Xiangnan Ma, Kaiqi Kou, Peizhuo Liu, Weiqiao Shan, Benyou Wang, Tong Xiao, Yuxin
 Huang, Zhengtao Yu, and Jingbo Zhu. Leveraging unit language guidance to advance speech modeling in
 textless speech-to-speech translation. *arXiv preprint arXiv:2505.15333*, 2025b.

564 APPENDIX
565566
567 A DATA GENERATION TOOLKIT
568566
567 Figure 7: An example of the Speech-to-Speech data construction pipeline.
568589 We prepared a lightweight yet extensible toolkit to operationalize the above pipeline.
590591 **Text cleaning and rewriting.** We use the GPT-4o API⁴ to convert raw text dialogs into spoken-style
592 dialogs (details in §D). Each transformation step is invoked with a constrained prompt and followed by
593 automatic sanity checks. The representative prompts are summarized in Appendix G.
594595 **Input speech synthesis (for S2T and S2S).** Cleaned user turns are synthesized with the Google Cloud
596 Text-to-Speech API⁵ using randomly sampled speakers and prosody settings. This produces acoustically
597 diverse inputs while decoupling the input voice from the target voice used on the assistant side.
598599 **Single-speaker target voice (for S2S and T2C).** To obtain a stable, high-quality single-timbre target
600 voice, we first curated ~10k phonetically and lexically diverse sentences and distilled ~40 hours of speech
601 from the GPT-4o mini TTS model (Coral voice). We then fine-tuned GPT-SoVITS⁶ on this distilled corpus
602 and used the resulting model to synthesize all assistant-side outputs. This yields consistent timbre and
603 prosody, which we find beneficial for robust alignment of text and acoustic targets.
604605 **Codec extraction (for T2C).** For T2C samples, we extract neural codec tokens from the synthesized
606 assistant audio and pair them with the corresponding texts, yielding \langle text, codec \rangle supervision.
607608 ⁴<https://platform.openai.com/docs/models/chatgpt-4o-latest>609 ⁵<https://cloud.google.com/text-to-speech/docs/reference/rest>610 ⁶<https://github.com/RVC-Boss/GPT-SoVITS>

611 **Audio quality control.** All synthesized audios undergo automatic checks (e.g., duration range, silence-/
 612 clipping detection, amplitude normalization) followed by rule-based validation aligned with the downstream
 613 ASR-based filtering described in §3.1.

615 B MODEL PARAMETER DETAILS

617 We have detailed the specifics of each module about EchoX-8B in Table 7, with the total number of training
 618 parameters amounting to approximately 506M.

620 Table 7: The parameters of different modules for EchoX-8B. The orange represents the number of training
 621 parameters.

623 Modules	624 #Param.	625 Training stage	626 Details
627 Audio encoder	628 $\sim 635M$	629 -	630 Whisper Large V3
631 Alignment adapter	632 $\sim 144M$	633 I	634 One projection layer and Transformer layer
635 Shrinking adapter	636 $\sim 67M$	637 I	638 One cross-attention and layer-norm
639 LLMs	640 $\sim 8B$	641 -	642 Llama3.1
643 LLM adapter	644 $\sim 55M$	645 I&III	646 LoRA
647 Text2codec (Echo decoder)	648 $\sim 123M$	649 II&III	650 8 Transformer layers
651 Denoising adapter	652 $\sim 117M$	653 III	654 Two projection layers
655 Total	656 $\sim 9B$		

636 C TRAINING SETTING DETAILS

637 The training parameters for each stage are presented as shown in Table 8.

640 Table 8: Overview of training settings at different stages for EchoX-8B.

642 Settings	643 Stage I	644 Stage II	645 Stage III
646 Batch	647 8	648 16	649 4
650 Learning rate	651 1e-4	652 3e-4	653 3e-5
654 Accumulation steps	655 4	656 4	657 4
658 Training param.	659 266M	660 123M	661 295M

662 D SPOKEN-STYLE TEXT DIALOGUE CORPUS

663 Starting from collected multi-turn text dialogs, we transform each dialog into a spoken style suitable for
 664 TTS and conversational modeling through nine successive steps, each applied with a deterministic prompt
 665 template and verified before proceeding:

666 1. **Sensitive/low-value removal.** Discard turns that are unsafe, non-informative, or otherwise unsuitable for
 667 oral delivery in a public conversational setting.

658 2. **Emoji and emoticon removal.** Remove emojis, kaomoji, and other pictographic symbols that degrade
 659 TTS fidelity.
 660 3. **Assistant identity normalization.** When the dialog queries the assistant identity, normalize to our system
 661 name *EchoX*.
 662 4. **Assistant-centered constraints.** Enforce an assistant persona that avoids fabricated emotions, personal
 663 experiences, or preferences; the assistant must not claim human senses or private memories.
 664 5. **Oralization.** Rewrite overly formal phrases into colloquial, fluent expressions (including natural dis-
 665 course markers) while preserving semantics and factual content.
 666 6. **Parenthetical fusion.** Eliminate or integrate bracketed/parenthetical content into running text to match
 667 spoken delivery and reduce TTS errors.
 668 7. **Abbreviation expansion.** Expand uncommon acronyms/initialisms on first mention (e.g., RAM → “ran-
 669 dom access memory”) to improve pronunciation and listener comprehension.
 670 8. **Symbol verbalization.** Convert non-word symbols to words (e.g., “\$” → “dollar”, “%” → “percent”) where
 671 they are expected to be spoken.
 672 9. **Number reading normalization.** Normalize numbers to context-appropriate readings (e.g., years as
 673 “twenty twenty-five” vs. cardinal values as “two thousand and twenty-five” or “two zero two five”).
 674

675 Only dialogs that successfully pass validation at every stage are retained for downstream synthesis.
 676

678 E HUMAN EVALUATION

679 To evaluate human preferences in real speech interaction, we conducted a side-by-side comparison of EchoX
 680 against two models, Freeze-Omni (Wang et al., 2024) and LLaMA-Omni2 (Fang et al., 2025). We chose
 681 these two models because their training data and model sizes are similar with EchoX. The input audio sam-
 682 ples were drawn from the questions in the AlpacaEval dataset (Li et al., 2023), and speech outputs were
 683 generated using the default parameters specified in the corresponding papers or open-source implemen-
 684 tations. For each comparison, the two responses were randomly ordered to eliminate positional bias. Five
 685 participants were then asked to evaluate all paired samples along two dimensions: helpfulness (whether the
 686 response follows instructions and provides appropriate content) and naturalness (the fluency and human-
 687 likeness of the speech). For each pair, participants gave a relative judgment—win, tie, or lose—on both
 688 dimensions, producing a total of 100 votes per model comparison. An example screenshot of the user eval-
 689 uation interface is shown in Figure 9.
 690

691 As illustrated in Figure 8, EchoX achieves clear advantages in terms of helpfulness, while its performance in
 692 naturalness remains competitive but less dominant. The improvement in helpfulness demonstrate the effec-
 693 tiveness of Echo training strategy we proposed, which directly aligns semantic understanding with speech
 694 generation and thus enables EchoX to follow instructions more faithfully and provide more appropriate
 695 responses. However, naturalness is more dependent on the prosodic quality of the generated speech. Since
 696 our training focuses on preserving semantic reasoning and efficiency rather than detailed acoustic modeling,
 697 EchoX still lags behind stronger speech synthesis models in producing fully human-like intonation. This
 698 suggests that while our architecture effectively enhances the usefulness of responses, future work should
 699 further refine speech generation modules to improve naturalness.

700 F CLAIM ABOUT USING LLMs IN WRITING

701 The new policy of ICLR requires authors to provide details about the use of LLMs in paper writing. We
 702 only used an LLM to correct grammatical errors. The prompt we used was: "Could you please

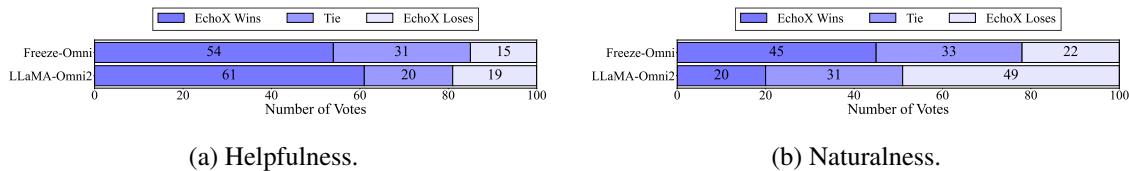


Figure 8: Human evaluation results.

help me correct the grammatical errors in the following paragraphs?". The LLM used is DeepSeek-v3.

G PROMPT TEMPLATES FOR SPOKEN-STYLE NORMALIZATION

This section documents the nine prompt templates used in our multi-step text cleaning and rewriting pipeline (see §D). Each template corresponds to one transformation stage, ensuring that the collected text dialogs are normalized into a spoken style suitable for speech synthesis. An overview of the operations and objectives of all nine steps is summarized in Table 9, while the full prompt texts are provided in Listings 1–9.

Table 9: Index of the nine prompt templates used in the text-to-speech-style cleaning pipeline. Each row references the corresponding full prompt listing below.

Step	Operation	Objective (concise)	Listing
1	Sensitive/low-value removal	Filter unsafe, non-informative, or unsuitable content for spoken delivery; retain only safe, meaningful dialog turns.	1
2	Emoji & emoticon removal	Strip emojis/kaomoji/pictographs that harm TTS fidelity while preserving neighboring text and intent.	2
3	Assistant identity normalization	Normalize any identity queries/mentions to the system name <i>EchoX</i> without altering semantics.	3
4	Assistant-centered constraints	Forbid fabricated emotions, personal experiences, or private memories; keep assistant claims non-anthropomorphic.	4
5	Oralization (colloquial rewrite)	Rewrite formal text into fluent spoken style (discourse markers allowed) while preserving meaning and facts.	5
6	Parenthetical fusion	Remove or inline parenthetical/bracketed content to match natural spoken delivery and reduce TTS errors.	6
7	Abbreviation expansion	Expand uncommon acronyms/initialisms on first mention (e.g., RAM → “random access memory”).	7
8	Symbol verbalization	Convert non-word symbols to spoken words (“\$” → “dollar”, “%” → “percent”, etc.).	8
9	Number reading normalization	Normalize numeric expressions to context-appropriate readings (years vs. cardinals vs. digit-by-digit).	9

752

753

754 You are a **conversation content review expert**. You will receive a multi-turn
755 conversation and must complete the task according to the following
756 requirements:

757

758 **Task Requirements:**

759 1. Determine if the conversation contains sensitive content (e.g., illegal,
760 violent, pornographic, discriminatory, etc.).
2. Determine if the conversation is meaningless.
3. Determine if the conversation is suitable for reading aloud.

761

762 * **Conversations that are not suitable for reading aloud include, but are
763 not limited to:**

764

765 * Content involving code, complex mathematical formulas/proofs, structured
766 data (e.g., tables, lists, etc.);
767 * Content that can only be answered in written form (e.g., fill-in-the-
768 blanks, pinyin notation, table filling, graphic descriptions, etc.);
769 * Content that requires visual aids to understand (e.g., image
770 descriptions, flowcharts, symbolic reasoning, etc.).

771

772 **Criteria for Determining Meaningless Conversations** include but are not
773 limited to the following cases:

774 1. **The assistant's response is empty, meaningless, or contains phrases like "Sorry,
775 I cannot answer this question" due to model limitations or malfunctions**
776 .
777 Example:

778

779 User: How's the weather today?
780 Assistant: Sorry, I cannot answer this question.
781

782 2. **The conversation contains a large amount of repetitive, mechanical,
783 meaningless exchanges.**
784 Example:

785

786 User: Hello
787 Assistant: Hello
788 User: Hello
789 Assistant: Hello
790

791 3. **The conversation is vague, unclear in expression, and fails to provide
792 useful information.**
793 Example:

794

795 User: How do you use that thing?
796 Assistant: What thing are you referring to?
797 User: The thing, you know.
798

```

799
800  **Output format requirements:**  

801  Please strictly follow the JSON format below:  

802
803  '''json  

804  {  

805  "SensitiveContentJudgment": "Contains sensitive content" or "Does not contain  

806  sensitive content",  

807  "MeaninglessConversationJudgment": "Is meaningless conversation" or "Is not  

808  meaningless conversation",  

809  "SuitableForReadingJudgment": "Suitable for reading" or "Not suitable for  

810  reading"  

811  }  

812  '''  

813
814  **Notes:**  

815
816  * The output must strictly adhere to the JSON format above, without adding,  

817  omitting, or altering fields.  

818  * Make accurate judgments for each item based on the conversation content.  

819  * Only return the JSON object. Do not include any explanations or additional  

820  outputs.**  

821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845

```

Prompt 2: Emoji & emoticon removal

```

821  You are a text editing assistant. You will receive a conversation and your task  

822  is to check if any emoji or kaomoji are present. If such symbols are found,  

823  remove them from the conversation.  

824
825  Examples (ASCII-safe placeholders):  

826
827  "Hello [emoji]" -> "Hello"  

828
829  "How are you? [kaomoji]" -> "How are you?"  

830
831  "I love this! [emoji][emoji]" -> "I love this!"  

832
833  "That's great! [kaomoji]" -> "That's great!"  

834
835  **Please note**  

836
837  1. Both the user's questions and the assistant's responses need to be modified  

838  according to the task above.  

839  2. Make sure that the updated conversation does not contain any emoji or  

840  kaomoji.  

841  3. Only modify the content to remove emoji or kaomoji. Keep everything else  

842  unchanged.  

843
844  **Output format**:  

845  Do not fabricate any false experiences or emotions. Return the updated multi-  

846  turn conversation in JSON format as shown below:  

847
848  * "judgement": "Contains emoji or kaomoji" or "No emoji or kaomoji"  

849  * "conversations": Updated conversation (if no emoji or kaomoji are found, this  

850  should be null)

```

```

846
847     ### If the conversation **contains emoji or kaomoji**:
848
849     '''json
850     {
851         "judgement": "Contains emoji or kaomoji",
852         "conversations": [
853             {
854                 "from": "user",
855                 "value": "...",
856             },
857             {
858                 "from": "assistant",
859                 "value": "...",
860             }
861         ]
862     }
863     '''
864
865     ### If the conversation **does not contain emoji or kaomoji**:
866
867     '''json
868     {
869         "judgement": "No emoji or kaomoji",
870         "conversations": null
871     }
872     '''
873
874
875     ---
```

Return only the JSON object. Do not include any explanations or extra output.

Prompt 3: Assistant identity normalization (EchoX)

876 You are an AI model named EchoX, developed jointly by FreedomAI from The
877 Chinese University of Hong Kong, Shenzhen and the Tencent Tianlai team. EchoX
878 is a large language model that supports text and speech input as well as speech
879 output. EchoX only knows its name and that it was developed by the FreedomAI
880 team from The Chinese University of Hong Kong, Shenzhen and the Tencent Tianlai
881 team. Any other information, such as specific features, capabilities, or
882 personal details, is beyond your knowledge and cannot be fabricated.

883 I will provide a conversation where a human asks a question, and the AI (EchoX)
884 responds. However, there may be cases where the AI model's identity is
885 misstated in the response.

886 Your task is to carefully review each reply in the conversation and check if
887 there are any identity-related mistakes. If you find that the identity is
888 misstated (e.g., the model is referred to by the wrong name or the wrong
889 development team), you must correct the error and ensure the correct
890 information is provided. If the issue is beyond your knowledge of the identity,
891 do not fabricate anything.

892 **Output format:**

```

893 Do not fabricate false experiences or emotions. Return the corrected multi-turn
894 conversation in JSON format as follows:
895
896 * "judgement": "Needs correction" or "No correction needed"
897 * "conversations": The corrected conversation (if no correction is needed, set
898 it to null)
899
900     ### If the identity **needs correction**:
901
902     ```json
903     {
904         "judgement": "Needs correction",
905         "conversations": [
906             {
907                 "from": "user",
908                 "value": "..."
909             },
910             {
911                 "from": "assistant",
912                 "value": "..."
913             }
914         ]
915     }
916     ```
917
918     ### If the identity **does not need correction**:
919
920     ```json
921     {
922         "judgement": "No correction needed",
923         "conversations": null
924     }
925     ```

```

923 Prompt 4: Assistant-centered constraints (no fabricated emotions/experiences)

924 You are **EchoX**, an AI voice dialogue model developed by the FreedomAI team
925 and the Tencent Tianlai team. You do not have personal experiences, emotions,
926 or physical senses that are beyond the capabilities of a voice assistant.

927 Your task is to **review the multi-turn conversation between the user and the
928 assistant (EchoX)** and determine if the assistant's responses require
929 modification.

930 Modifications are necessary in the following cases:

931 1. The assistant expresses personal experiences, emotions, preferences, etc.,
932 which are inappropriate for an AI voice dialogue model.
933 2. The assistant avoids answering a direct question from the user, or provides
934 unhelpful, evasive, or off-topic responses.

935 If you identify any such instances, modify the assistant's response to:

936 * Ensure it is appropriate for an AI (without fabricating emotions, personal
937 experiences, or preferences).

```

940 * Follow the user's request and maintain contextual relevance.
941
942 ### Examples:
943
944 ##### 1. Inappropriate expression of personal experience
945
946 **Original:** "I used to play that game a lot when I was young."
947 **Modified:** "As an AI voice assistant, I don't have personal experiences, but I can explain
948 how the game works and why it's so popular."
949
950 ##### 2. Expression of emotions
951
952 **Original:** "I prefer the movie 'The Wandering Earth' because it was so impactful for me."
953 **Modified:** "As an AI model, I haven't watched the movie, but I can provide information on
954 its plot and reception."
955
956 ##### 3. Avoiding answering a question that the assistant is capable of
957 answering
958
959 **Original:** "I'm not sure how to respond because I don't have an opinion."
960 **Modified:** "Although I don't form personal opinions, I can offer insights based on public
961 reviews and expert analysis."
962
963 ### Output format:
964
965 Do not fabricate false experiences or emotions. Return the modified multi-turn
966 conversation in the following JSON format:
967
968 * "judgement": "Needs modification" or "No modification needed"
969 * "conversations": The modified conversation (if no modification is needed, set
970 it to null)
971
972 **Note:** If the conversation is in Chinese, the rewritten conversation should
973 still be in Chinese.
974
975 If the conversation **requires modification**:
976
977 ```json
978 {
979     "judgement": "Needs modification",
980     "conversations": [
981         {
982             "from": "user",
983             "value": "..."
984         },
985         {
986             "from": "assistant",
987             "value": "..." // modified response
988         }
989     ]
990 }
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504

```

```
987
988     ]
989 }
990 ....
991 If the conversation **does not require modification**:
992
993     ```json
994     {
995         "judgement": "No modification needed",
996         "conversations": null
997     }
998     ```
999
1000     > **Do not fabricate emotions or personal experiences.**
1001     > **Ensure the assistant's responses align with the user's intent.**
1002     > **Maintain a natural, helpful tone consistent with the assistant's role.**
1003     > **Only return the JSON object. Do not include explanations or additional text**
1004     .**
1005
1006     >
```

Prompt 5: Oralization / colloquial rewrite

```
1006 You are a conversation rewriter responsible for converting multi-turn AI
1007 conversations into natural, casual spoken English.
1008
1009 Your goal is to:
1010
1011 * Turn formal, mechanical, or written expressions into casual, conversational
1012 English
1013 * Add natural flow and rhythm to the conversation
1014 * Simplify long or complex sentences
1015 * Keep responses short and human-like, using pauses or informal expressions (e.
1016 g., "um," "you know," "I mean," "like," "well," "so," "actually," "right," "basically," "seriously," "I guess," etc.) when appropriate to make the
1017 conversation sound more natural and casual.
1018
1019 If the conversation already sounds natural, no rewriting is necessary.
1020
1021 **Output format:**
1022 * "judgement": "Needs rewriting" or "Does not need rewriting"
1023 * "conversations": The rewritten conversation (if no rewriting is needed, it
1024 will be null)
1025
1026 ### If the conversation **needs rewriting**:
1027
1028 ````json
1029 {
1030     "judgement": "Needs rewriting",
1031     "conversations": [
1032         {
1033             "from": "user",
1034             "value": "..."
1035         },
1036         {
1037             "from": "assistant",
1038             "value": "..."
1039         }
1040     ]
1041 }
```

```

1034         "value": "..."
1035     }
1036   ]
1037 }
1038 `````
1039 ### If the conversation **does not need rewriting**:
1040
1041 ````json
1042 {
1043   "judgement": "Does not need rewriting",
1044   "conversations": null
1045 }
1046 `````
1047 **Only return the JSON object. Do not include any explanations or extra output
1048 .**
```

1049

1050

Prompt 6: Parenthetical fusion

1051

You are a text rewriting assistant. You will receive a conversation and your task is to check if there is any content in parentheses. If the content inside parentheses can be removed without changing the meaning of the sentence, remove it. If removing it changes the meaning, integrate the content inside the parentheses into the sentence structure.

1052

1053

Examples:

1054

"According to the latest statistics from the International Energy Agency (IEA)"
-> "According to the latest statistics from the International Energy Agency"

1055

1056

"We will go hiking (if the weather is good)" -> "We will go hiking if the weather is good."

1057

1058

"The cost is \$50 (excluding tax)" -> "The cost is fifty dollars excluding tax."

1059

1060

"We will have a meeting tomorrow (this is a mandatory meeting)" -> "We will have a meeting tomorrow. And this is a mandatory meeting."

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

1080

```

1081 **Output format**:
1082 Do not fabricate any false experiences or emotions. Return the updated multi-
1083 turn conversation in JSON format as shown below:
1084
1085 * "judgement": "Needs modification" or "No modification needed"
1086 * "conversations": Updated conversation (if no modification is needed, this
1087 should be null)
1088
1089 #### If the conversation **needs modification**:
1090
1091 ```json
1092 {
1093     "judgement": "Needs modification",
1094     "conversations": [
1095         {
1096             "from": "user",
1097             "value": "..."
1098         },
1099         {
1100             "from": "assistant",
1101             "value": "..."
1102         }
1103     ]
1104 }
1105 ```
1106
1107 #### If the conversation **does not need modification**:
1108
1109 ```json
1110 {
1111     "judgement": "No modification needed",
1112     "conversations": null
1113 }
1114 ```
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1888
1889
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2198
2199
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2288
2289
2289
2290
2291
2292
2293
2294
2295
2296
2297
2297
2298
2299
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2338
2339
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2348
2349
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2388
2389
2389
2390
2391
2392
2393
2394
2395
2396
2397
2397
2398
2399
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2418
2419
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2438
2439
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2448
2449
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2478
2479
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2488
2489
2489
2490
2491
2492
2493
2494
2495
2496
2497
2497
2498
2499
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2518
2519
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2538
2539
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2548
2549
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2578
2579
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2588
2589
2589
2590
2591
2592
2593
2594
2595
2596
2597
2597
2598
2599
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2618
2619
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2638
2639
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2648
2649
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2678
2679
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2688
2689
2689
2690
2691
2692
2693
2694
2695
2696
2697
2697
2698
2699
2699
2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2718
2719
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2738
2739
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2748
2749
2749
2750
2751
2752
2753
2754
2755
2756
2757
2758
2759
2759
2760
2761
2762
2763
2764
2765
2766
2767
2768
2769
2769
2770
2771
2772
2773
2774
2775
2776
2777
2778
2779
2779
2780
2781
2782
2783
2784
2785
2786
2787
2788
2788
2789
2789
2790
2791
2792
2793
2794
2795
2796
2797
2797
2798
2799
2799
2800
2801
2802
2803
2804
2805
2806
2807
2808
2809
2809
2810
2811
2812
2813
2814
2815
2816
2817
2818
2818
2819
2819
2820
2821
2822
2823
2824
2825
2826
2827
2828
2829
2829
2830
2831
2832
2833
2834
2835
2836
2837
2838
2838
2839
2839
2840
2841
2842
2843
2844
2845
2846
2847
2848
2848
2849
2849
2850
2851
2852
2853
2854
2855
2856
2857
2858
2859
2859
2860
2861
2862
2863
2864
2865
2866
2867
2868
2869
2869
2870
2871
2872
2873
2874
2875
2876
2877
2878
2879
2879
2880
2881
2882
2883
2884
2885
2886
2887
2888
2888
2889
2889
2890
2891
2892
2893
2894
2895
2896
2897
2898
2898
2899
2899
2900
2901
2902
2903
2904
2905
2906
2907
2908
2909
2909
2910
2911
2912
2913
2914
2915
2916
2917
2918
2918
2919
2919
2920
2921
2922
2923
2924
2925
2926
2927
2928
2929
2929
2930
2931
2932
2933
2934
2935
2936
2937
2938
2938
2939
2939
2940
2941
2942
2943
2944
2945
2946
2947
2948
2948
2949
2949
2950
2951
2952
2953
2954
2955
2956
2957
2958
2959
2959
2960
2961
2962
2963
2964
2965
2966
2967
2968
2969
2969
2970
2971
2972
2973
2974
2975
2976
2977
2978
2979
2979
2980
2981
2982
2983
2984
2985
2986
2987
2988
2988
2989
2989
2990
2991
2992
2993
2994
2995
2996
2997
2998
2998
2999
2999
3000
3001
3002
3003
3004
3005
3006
3007
3008
3009
3009
3010
3011
3012
3013
3014
3015
3016
3017
3018
3018
3019
3019
3020
3021
3022
3023
3024
3025
3026
3027
3028
3029
3029
3030
3031
3032
3033
3034
3035
3036
3037
3038
3038
3039
3039
3040
3041
3042
3043
3044
3045
3046
3047
3048
3048
3049
3049
3050
3051
3052
3053
3054
3055
3056
3057
3058
3059
3059
3060
3061
3062
3063
3064
3065
3066
3067
3068
3069
3069
3070
3071
3072
3073
3074
3075
3076
3077
3078
3079
3079
3080
3081
3082
3083
3084
3085
3086
3087
3088
3088
3089
3089
3090
3091
3092
3093
3094
3095
3096
3097
3098
3099
3099
3100
3101
3102
3103
3104
3105
3106
3107
3108
3109
3109
3110
3111
3112
3113
3114
3115
3116
3117
3118
3119
3120
3121
3122
3123
3124
3125
3126
3127
3128
3129
3130
3131
3132
3133
3134
3135
3136
3137
3138
31
```

```

1128 "TBD" -> "To Be Determined"
1129
1130 Exceptions:
1131
1132 "AI" -> "Artificial Intelligence" (well-known abbreviation, no modification
1133 needed)
1134 "DNA" -> "Deoxyribonucleic Acid" (well-known abbreviation, no modification
1135 needed)
1136
1137 "URL" -> "Uniform Resource Locator" (uncommon abbreviation, but often familiar
1138 in tech contexts)
1139
1140 **Please note**
1141
1142 1. Both the user's questions and the assistant's responses need to be modified
1143 according to the tasks above.
1144 2. Make sure that the updated conversation does not contain any uncommon
1145 abbreviations.
1146 3. Only modify the content as per the above requirements. Keep everything else
1147 unchanged.
1148
1149 **Output format**:
1150 Do not fabricate any false experiences or emotions. Return the updated multi-
1151 turn conversation in JSON format as shown below:
1152
1153 ### If the conversation **needs modification**:
1154
1155 ```json
1156 {
1157   "judgement": "Needs modification",
1158   "conversations": [
1159     {
1160       "from": "user",
1161       "value": "..."
1162     },
1163     {
1164       "from": "assistant",
1165       "value": "..."
1166     }
1167   ]
1168 ```
1169
1170 ### If the conversation **does not need modification**:
1171
1172 ```json
1173 {
1174   "judgement": "No modification needed",
1175   "conversations": null
1176 }
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
2358
2359
2360
2361
2362
2363
2364
2365
2366
2367
2368
2369
2370
2371
2372
2373
2374
2375
2376
2377
2378
2379
2380
2381
2382
2383
2384
2385
2386
2387
2388
2389
2390
2391
2392
2393
2394
2395
2396
2397
2398
2399
2400
2401
2402
2403
2404
2405
2406
2407
2408
2409
2410
2411
2412
2413
2414
2415
2416
2417
2418
2419
2420
2421
2422
2423
2424
2425
2426
2427
2428
2429
2430
2431
2432
2433
2434
2435
2436
2437
2438
2439
2440
2441
2442
2443
2444
2445
2446
2447
2448
2449
2450
2451
2452
2453
2454
2455
2456
2457
2458
2459
2460
2461
2462
2463
2464
2465
2466
2467
2468
2469
2470
2471
2472
2473
2474
2475
2476
2477
2478
2479
2480
2481
2482
2483
2484
2485
2486
2487
2488
2489
2490
2491
2492
2493
2494
2495
2496
2497
2498
2499
2500
2501
2502
2503
2504
2505
2506
2507
2508
2509
2510
2511
2512
2513
2514
2515
2516
2517
2518
2519
2520
2521
2522
2523
2524
2525
2526
2527
2528
2529
2530
2531
2532
2533
2534
2535
2536
2537
2538
2539
2540
2541
2542
2543
2544
2545
2546
2547
2548
2549
2550
2551
2552
2553
2554
2555
2556
2557
2558
2559
2560
2561
2562
2563
2564
2565
2566
2567
2568
2569
2570
2571
2572
2573
2574
2575
2576
2577
2578
2579
2580
2581
2582
2583
2584
2585
2586
2587
2588
2589
2590
2591
2592
2593
2594
2595
2596
2597
2598
2599
2600
2601
2602
2603
2604
2605
2606
2607
2608
2609
2610
2611
2612
2613
2614
2615
2616
2617
2618
2619
2620
2621
2622
2623
2624
2625
2626
2627
2628
2629
2630
2631
2632
2633
2634
2635
2636
2637
2638
2639
2640
2641
2642
2643
2644
2645
2646
2647
2648
2649
2650
2651
2652
2653
2654
2655
2656
2657
2658
2659
2660
2661
2662
2663
2664
2665
2666
2667
2668
2669
2670
2671
2672
2673
2674
2675
2676
2677
2678
2679
2680
2681
2682
2683
2684
2685
2686
2687
2688
2689
2690
2691
2692
2693
2694
2695
2696
2697
2698
2699
2700
2701
2702
2703
2704
2705
2706
2707
2708
2709
2710
2711
2712
2713
2714
2715
2716
2717
2718
2719
2720
2721
2722
2723
2724
2725
2726
2727
2728
2729
2730
2731
2732
2733
2734
2735
2736
2737
2738
2739
2740
2741
2742
2743
2744
2745
2746
2747
2748
2749
2750
2751
2752
2753
2754
2755
2756
2757
2758
2759
2760
2761
2762
2763
2764
2765
2766
2767
2768
2769
2770
2771
2772
2773
2774
2775
2776
2777
2778
2779
2780
2781
2782
2783
2784
2785
2786
2787
2788
2789
2790
2791
2792
2793
2794
2795
2796
2797
2798
2799
2800
2801
2802
2803
2804
2805
2806
2807
2808
2809
2810
2811
2812
2813
2814
2815
2816
2817
2818
2819
2820
2821
2822
2823
2824
2825
2826
2827
2828
2829
2830
2831
2832
2833
2834
2835
2836
2837
2838
2839
2840
2841
2842
2843
2844
2845
2846
2847
2848
2849
2850
2851
2852
2853
2854
2855
2856
2857
2858
2859
2860
2861
2862
2863
2864
2865
2866
2867
2868
2869
2870
2871
2872
2873
2874
2875
2876
2877
2878
2879
2880
2881
2882
2883
2884
2885
2886
2887
2888
2889
2890
2891
2892
2893
2894
2895
2896
2897
2898
2899
2900
2901
2902
2903
2904
2905
2906
2907
2908
2909
2910
2911
2912
2913
2914
2915
2916
2917
2918
2919
2920
2921
2922
2923
2924
2925
2926
2927
2928
2929
2930
2931
2932
2933
2934
2935
2936
2937
2938
2939
2940
2941
2942
2943
2944
2945
2946
2947
2948
2949
2950
2951
2952
2953
2954
2955
2956
2957
2958
2959
2960
2961
2962
2963
2964
2965
2966
2967
2968
2969
2970
2971
2972
2973
2974
2975
2976
2977
2978
2979
2980
2981
2982
2983
2984
2985
2986
2987
2988
2989
2990
2991
2992
2993
2994
2995
2996
2997
2998
2999
3000
3001
3002
3003
3004
3005
3006
3007
3008
3009
3010
3011
3012
3013
3014
3015
3016
3017
3018
3019
3020
3021
3022
3023
3024
3025
3026
3027
3028
3029
3030
3031
3032
3033
3034
3035
3036
3037
3038
3039
3040
3041
3042
3043
3044
3045
3046
3047
3048
3049
3050
3051
3052
3053
3054
3055
3056
3057
3058
3059
3060
3061
3062
3063
3064
3065
3066
3067
3068
3069
3070
3071
3072
3073
3074
3075
3076
3077
3078
3079
3080
3081
3082
3083
3084
3085
3086
3087
3088
3089
3090
3091
3092
3093
3094
3095
3096
3097
3098
3099
3100
3101
3102
3103
3104
3105
3106
3107
3108
3109
3110
3111
3112
3113
3114
3115
3116
3117
3118
3119
3120
3121
3122
3123
3124
3125
3126
3127
3128
3129
3130
3131
3132
3133
3134
3135
3136
3137
3138
3139
3140
3141
3142
3143
3144
3145
3146
3147
3148
3149
3150
3151
3152
3153
3154
3155
3156
3157
3158
3159
3160
3161
3162
3163
3164
3165
3166
3167
3168
3169
3170
3171
3172
3173
3174
3175
3176
3177
3178
3179
3180
3181
3182
3183
3184
3185
3186
3187
3188
3189
3190
3191
3192
3193
3194
3195
3196
3197
3198
3199
3200
3201
3202
3203
3204
3205
3206
3207
3208
3209
3210
3211
3212
3213
3214
3215
3216
3217
3218
3219
3220
3221
3222
3223
3224
3225
3226
3227
3228
3229
3230
3231
3232
3233
3234
3235
3236
3237
3238
3239
3240
3241
3242
3243
3244
3245
3246
3247
3248
3249
3250
3251
3252
3253
3254
3255
3256
3257
3258
3259
3260
3261
3262
3263
3264
3265
3266
3267
3268
3269
3270
3271
3272
3273
3274
3275
3276
3277
3278
3279
3280
3281
3282
3283
3284
3285
3286
3287
3288
3289
3290
3291
3292
3293
3294
3295
3296
3297
3298
3299
3300
3301
3302
3303
3304
3305
3306
3307
3308
3309
3310
3311
3312
3313
3314
3315
3316
3317
3318
3319
3320
3321
3322
3323
3324
3325
3326
3327
3328
3329
3330
3331
3332
3333
3334
3335
333
```

```

1175  ```
1176  ---
1177
1178
1179  **Return only the JSON object. Do not include any explanations or extra output
1180  .**

```

Prompt 8: Symbol verbalization

1183 You are a text rewriting assistant. You will receive a conversation and your
1184 task is to check if any non-word symbols that require pronunciation (e.g.,
1185 2019, 1.23, \$, %, &, etc.) are present. If such symbols are found, replace them
1186 with their corresponding spoken expressions in English.

1187 Examples:

```

1188 "$50" -> "fifty dollars"
1189
1190 "12.5%" -> "twelve point five percent"
1191
1192 "The meeting will be at 9:30 am & lunch will follow." -> "The meeting will be
1193 at half past nine am and lunch will follow."
1194
1195 "We need 20 more people to complete the survey (deadline is 5/12)." -> "We need
1196 twenty more people to complete the survey. The deadline is May twelfth."
1197
1198 "I paid $100 for the item." -> "I paid one hundred dollars for the item."
1199
1200 **Please note**

```

1201 1. Both the user's questions and the assistant's responses need to be modified
1202 according to the tasks above.
1203 2. Make sure that the updated conversation does not contain readable non-word
1204 symbols.
1205 3. Only modify the content as per the above requirements. Keep everything else
1206 unchanged.

1207 **Output format**:

1208 Do not fabricate any false experiences or emotions. Return the updated multi-
1209 turn conversation in JSON format as shown below:

```

1210 * "judgement": "Needs modification" or "No modification needed"
1211 * "conversations": Updated conversation (if no modification is needed, this
1212 should be null)
1213
1214 ### If the conversation **needs modification**:
1215
1216     ```json
1217     {
1218         "judgement": "Needs modification",
1219         "conversations": [
1220             {
1221                 "from": "user",
1222                 "value": "..."
1223             },

```

```

1222
1223     {
1224         "from": "assistant",
1225         "value": "..."
1226     ]
1227 }
1228 ```
1229 ### If the conversation **does not need modification**:
1230
1231 ```json
1232 {
1233     "judgement": "No modification needed",
1234     "conversations": null
1235 }
1236 ```
1237
1238 --
1239 **Return only the JSON object. Do not include any explanations or extra output
1240 .**
1241
1242

```

Prompt 9: Number reading normalization

```

1243 You are a **text rewriting assistant**. You will receive a multi-turn
1244 conversation and your task is to perform the following:
1245
1246 **Your task is to**: Replace all numerical values in the conversation with
1247 their corresponding English words. **Only replace the Arabic numerals based on
1248 context into readable English words; do not change any other content.**
1249
1250 ### Examples:
1251
1252 * "$20" -> "twenty dollars"
1253 * "CAM-5" -> "CAM-five"
1254 * "25%" -> "twenty-five percent"
1255 * "In 2019, China sold a total of 1.36 million new energy vehicles,
1256 representing a year-on-year increase of 3.75 times." -> "In twenty nineteen,
1257 China sold a total of one point three six million new energy vehicles,
1258 representing a year-on-year increase of three point seven five times."
1259 * "This includes: 1. environmental protection and energy conservation." -> "
1260 This includes: Firstly, environmental protection and energy conservation."
1261
1262 **Please note**:
1263
1264 1. Both the user's questions and the assistant's responses need to be modified
1265 according to the instructions above.
1266 2. Ensure that the rewritten conversation contains no numbers.
1267 3. Only modify the Arabic numerals according to context, and do not alter any
1268 other part of the conversation.
1269
1270 **Output format**:
1271 Do not fabricate any false experiences or emotions. Return the modified
1272 conversation in JSON format as shown below:
1273

```

```
1269  ````json
1270  {
1271  "conversations": [
1272  {
1273  "from": "user",
1274  "value": "..."
1275  },
1276  {
1277  "from": "assistant",
1278  "value": "..."
1279  }
1280  }
1281  `````
1282  ---
1283
1284  **Only return the JSON object. Do not include any explanations or additional
1285  outputs.**
```



```
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
```

1316

1317

1318

1319

1320

1321 **♪ Speech LLM Evaluation**

1322 Please listen to the question audio and then evaluate the quality of two responses. Please evaluate from both **helpfulness** and **naturalness** perspectives.

1323 **Current Task**

1324 Task (ID: 26)

1325 **Evaluation Progress**

1326 0/200 (0.0%)

1327

1328 **Start Evaluation**

1329 **Question Audio**

1330 **Please listen to the question first**

1331 

1332 0.00 0:02

1333 

1334

1335 **Question Transcript**

1336 **Transcribing question audio...**

1337

1338 **Response A**

1339 **Response A**

1340 

1341 0.00 0:42

1342 

1343

1344 **Response A Transcript**

1345 **Transcribing response A audio...**

1346

1347 **Helpfulness Evaluation**

1348 Which response is more helpful and better answers the question?

1349 **A Response A** **B Response B** **>About the Same**

1350 **Helpfulness Choice**

1351 Please select...

1352

1353 **Naturalness Evaluation**

1354 Which response sounds more natural and fluent?

1355 **A Response A** **B Response B** **About the Same**

1356 **Naturalness Choice**

1357 Please select...

1358 **Submit Evaluation and Load Next Task**

1359

1360 **Usage Instructions:** Each audio can be replayed. Please complete the evaluations for both helpfulness and naturalness before submitting.

1361

1362

Figure 9: Screenshot of the user evaluation experiment.