

# InSertter: Speech Instruction Following with Unsupervised Interleaved Pre-training

Dingdong Wang<sup>1\*</sup>, Jin Xu<sup>2\*</sup>, Ruihang Chu<sup>1</sup>, Zhifang Guo, Xiong Wang, Jincenzi Wu<sup>1</sup>  
Dongchao Yang<sup>1</sup>, Shengpeng Ji, Junyang Lin<sup>2</sup>

<sup>1</sup>The Chinese University of Hong Kong, <sup>2</sup>Alibaba Group  
dingdongwang@link.cuhk.edu.hk

## Abstract

Recent advancements in speech large language models (SpeechLLMs) have attracted considerable attention. Nonetheless, current methods exhibit suboptimal performance in adhering to speech instructions. Notably, the intelligence of models significantly diminishes when processing speech-form input as compared to direct text-form input. Prior work has attempted to mitigate this semantic inconsistency between speech and text representations through techniques such as representation and behavior alignment, which involve the meticulous design of data pairs during the post-training phase. In this paper, we introduce a simple and scalable training method called **InSertter**, which stands for **I**nterleaved **S**peech-**T**ext **R**epresentation **P**re-training. InSertter is designed to pre-train large-scale unsupervised speech-text sequences, where the speech is synthesized from randomly selected segments of an extensive text corpus using text-to-speech conversion. Consequently, the model acquires the ability to generate textual continuations corresponding to the provided speech segments, obviating the need for intensive data design endeavors. To systematically evaluate speech instruction-following capabilities, we introduce SpeechInstructBench, the first comprehensive benchmark specifically designed for speech-oriented instruction-following tasks. Our proposed model InSertter achieves SOTA performance in SpeechInstructBench and demonstrates superior or competitive results across diverse speech processing tasks. SpeechInstructBench is available at <https://huggingface.co/datasets/ddwang2000/SpeechInstructBench>.

## 1 Introduction

Recently, speech large language models (SpeechLLMs) have garnered significant attention, as they

\* Equal contribution.

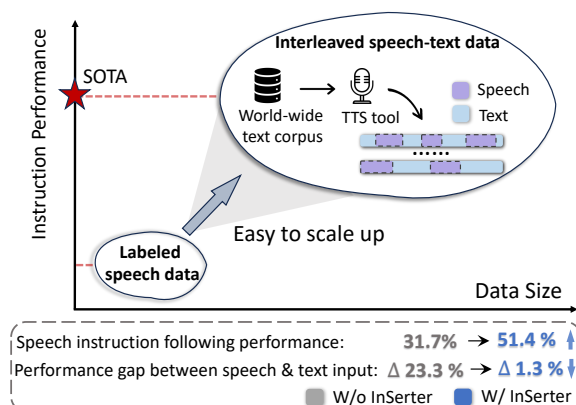


Figure 1: InSertter leverages large-scale interleaved data for pre-training. It improves instruction-following accuracy (SpeechInstructBench closed-ended standard version) given speech-form input from 31.7% to 51.4%, further narrowing the gap with text-form inputs from 23.3% to 1.3% compared to non-interleaved baseline.

hold promise for enabling more natural and intuitive human-machine interactions (Ji et al., 2024; Chu et al., 2023, 2024; Défossez et al., 2024; Tang et al., 2024). SpeechLLMs receive speech inputs and directly follow speech-based instructions to perform responses. However, due to the intrinsic differences between speech and text representations, it is challenging for models to effectively comprehend the consistency in semantic content expression across different modalities. This discrepancy leads to difficulties in accurately following spoken instructions compared to text-based ones.

Enabling the speech modality to inherit the cognitive capabilities of text represents a fundamental challenge in the domain of end-to-end speech instruction following. Prior research has attempted to address this challenge through two approaches: representation alignment (Held et al., 2024) and behavior alignment (Wang et al., 2024; Held et al., 2024). However, representation alignment encounters limitations stemming from the inherent differences in granularity, such as the varying sequence lengths

and the discrete nature of text tokens compared to continuous speech frames. Consequently, the direct alignment of these distinct representations often results in the degradation of critical original speech features, including tone, energy, and pitch. On the other hand, behavior alignment focuses on training models to learn consistent response sequences or distributions given either speech input or its transcripts during the post-training phase. Nonetheless, these approaches face significant scalability challenges due to their selected training stage, the complexity of data construction, and overall training efficiency. Therefore, there remains substantial scope for further improvement.

The emergence of intelligence in text is fundamentally based on unsupervised next-token prediction during the pre-training phase. Specifically, the training objective is to predict the subsequent continuation token based on a sequence of preceding text. Therefore, the motivation for our work is to enable speech sequences to inherit the intelligence characteristic of text by requiring the model to execute the following task: *given a speech sequence, predict the next text token or segment continuation during the pre-training phase*. By interleaving speech and text sequences in training samples, we allow for multi-turn text-speech alignment and thus boost training efficiency.

To this end, we propose **InSert**, which stands for **Interleaved Speech-Text Representation Pre-training**. InSert employs text-to-speech to synthesize speech segments from randomly selected portions of large-scale text pre-training corpora, thereby ensuring scalability. InSert provides several key advantages: 1) it facilitates dynamic interactions between speech and text modalities through an interleaved training format; 2) it enhances training efficiency by enabling simultaneous learning of multiple segment relationships during pre-training; and 3) it offers high scalability by leveraging existing text corpora via text-to-speech conversion.

To comprehensively evaluate models’ performance on speech instruction following, we introduce SpeechInstructBench. This benchmark is characterized by the incorporation of 1) linguistic variations in speech, which encompass spontaneous speech phenomena such as disfluencies and self-corrections, as well as diverse accents; 2) acoustic variations that include speaker-specific traits such as pitch modulation, speech rate, emotional tone, and energy levels; and 3) environmental variations, which account for background noise to sim-

ulate real-world interaction conditions. Our experiments show that InSert achieves state-of-the-art performance on SpeechInstructBench and significantly improves results across multiple benchmarks within VoiceBench (Chen et al., 2024b).

We summarize our contributions as follows:

- We propose InSert, a conceptually simple, powerful, and scalable pre-training algorithm to significantly improve speech instruction following capabilities of SpeechLLMs.
- We present SpeechInstructBench, the first comprehensive and bilingual (English-Chinese) benchmark designed to evaluate speech instruction-following abilities.
- InSert attains state-of-the-art performance on SpeechInstructBench and achieves significant improvement across various speech processing tasks in VoiceBench (Chen et al., 2024b).

## 2 Related Work

**SpeechLLMs.** Current SpeechLLMs are generally categorized by their speech representation strategies, using either discrete speech tokens or continuous speech features as input. The first category converts speech inputs into discrete tokens (Défossez et al., 2024; Chen et al., 2024a; Veluri et al., 2024; Zhang et al., 2023; Rubenstein et al., 2023; Borsos et al., 2023; Xie and Wu, 2024b), leveraging self-supervised speech pre-training to expand the vocabulary of LLMs with learned speech representations. Some studies (Défossez et al., 2024; Nguyen et al., 2024; Zeng et al., 2024b) incorporate synthetic interleaved data for pre-training. However, there are two notable distinctions between these works and InSert. First, the primary objective of these studies is to enhance the quality and end-to-end coherence of speech generation, while InSert aims to improve models’ speech instruction-following capabilities. Second, existing works rely on discrete speech sequence modeling, whereas InSert is specifically designed for models with continuous speech representations, such as Qwen-Audio series (Chu et al., 2023, 2024).

The second approach processes continuous speech signals directly (Chen et al., 2023; Gong et al., 2023a,b; Ma et al., 2024; Xue et al., 2024; Chu et al., 2023, 2024; Tang et al., 2024). This strategy avoids discretization loss and retains rich speech characteristics, such as intonation, rhythm, and speaker-specific attributes. Typically, a pre-



including removing special characters and standardization via regular expressions to improve the quality, giving rise to a filtered text corpus comprising around 610 billion text tokens. Detailed statistics of the corpus are provided in Appendix A.

**(ii) Segments sampling.** Given the source text corpus, we sample portions of text and convert them into speech segments. Our sampling strategy operates at two granularity levels: word-level and sentence-level. For word-level sampling, we randomly select certain words as the segments for conversion, maintaining a minimum length of five words to preserve semantic coherence. For example, regarding “[The weather is really nice] today, I want to go for [a walk in the park] in the afternoon”, the portions in brackets represent the speech segments we sampled, while the remaining are text segments. For sentence-level sampling, we use punctuation marks as delimiters to sample sentences as speech segments randomly.

**(iii) Text-to-speech conversion.** We convert the sampled speech segments by CosyVoice 2.0 model (Du et al., 2024) with multi-speaker voice prompts. Specifically, to enhance the diversity of the synthesized speech, we select 10,000 unique voice prompts from an extensive timbre library by evaluating Word Error Rate (WER) and WV-MOS (Andreev et al., 2023). The final generated speech corpus consists of 301,540 hours of speech data, together with the remaining text data to form the interleaved speech-text dataset.

### 3.2 Training Strategy

The prepared interleaved data is adopted to pre-train our SpeechLLM, which can significantly enhance the model’s ability to follow speech instructions through text-speech alignment. Then we conduct a supervised fine-tuning stage using dialogue data to enhance the model’s interactive performance. We detail the two-stage training below.

**Stage 1: Pre-training on interleaved data.** In-Serter guides the model to align speech embeddings with text tokens through next-token prediction. As shown in Fig. 2(b), we process speech segments with a speech encoder and an adapter, resulting in continuous speech representations and interleaved with text segments. Let  $S = (S_1, \dots, S_i, \dots, S_N)$  denote the sequence of speech embeddings from  $N$  speech segments, and  $Y = (Y_1, \dots, Y_i, \dots, Y_N)$  denote the corresponding text token sequences. Each text token segment  $Y_i$  consists of  $M_i$  tokens  $(Y_{i,1}, \dots, Y_{i,M_i})$ . The training objective is to

minimize the cross-entropy loss for predicting the next text token conditioned on both the preceding speech embeddings and text tokens:

$$L_{CE} = - \sum_{i=1}^N \sum_{j=1}^{M_i} \log p(Y_{i,j} \mid S_{1:i}, Y_{<i,j}; \theta),$$

where  $Y_{i,j}$  represents the predicted  $j$ -th text token in  $i$ -th text token segment. Each  $Y_{i,j}$  is predicted conditional on all preceding speech embeddings  $S_{1:i}$  (from the first to the  $i$ -th segment) and all previous text tokens  $Y_{<i,j}$  (up to the  $j$ -th text token in the  $i$ -th text token segment).  $\theta$  is all trainable model parameters, including those within the adapter (Q-Former) and the LLM. Importantly, as indicated in Fig. 2(b), cross-entropy loss is computed only for tokens corresponding to ground-truth text segments, while tokens associated with speech segments are masked out during loss calculation.

In addition to interleaved data, in the pre-training stage, we follow the same multi-task speech data categories (Chu et al., 2024, 2023) and collect the speech data to strengthen the bidirectional speech-text relationships. Text-only corpora are also added to preserve the model’s strong text generation quality. We balance these data in a proportion of 40% (interleaved), 30% (multi-task speech data), and 30% (text-only).

**Stage 2: Supervised fine-tuning.** After the pre-training stage aligns speech and text modalities, we conduct supervised fine-tuning (SFT) to enhance the model’s dialogue capabilities. We collect data from speech dialogue and text dialogue and empirically incorporate 50% text data during this stage. The total amount of SFT data is 20K samples.

## 4 SpeechInstructBench

Instruction-following capability refers to a model’s ability to precisely interpret and execute natural language directives from users while adhering to all stated constraints and requirements (Zhou et al., 2023). However, the current existing benchmarks on SpeechLLMs primarily evaluate response quality rather than instruction following accuracy. To bridge this gap, we present SpeechInstructBench, a comprehensive and bilingual (English-Chinese) speech instruction-following benchmark for SpeechLLMs. As shown in Table 1, this benchmark encompasses three evaluation sub-tasks: closed-ended instructions, open-ended instructions, and adjustment instructions.



Types	Tasks	Num	Instruction Examples
Closed-Ended	Standard		Include keywords {keyword} in your response at least three times.
	Accent	7718 (En)	You should not use {forbidden words} in the response.
	Background	5208 (Cn)	Entire output should be wrapped in {JSON} format.
	Disfluency		Your entire response should be in English, {capital letters only}.
Open-Ended	Paralanguage		Uh...finish response with um...the exact phrase {end phrase}. And... and... no other words should follow this phrase.
	Content		Give two different responses and should be separated by {6 continuous asterisk symbols}.
	Format	500 (En)	Recommend me {5 Chinese films released before 1990}.
	Style	500 (Cn)	Give me suggestions on keeping health, {use bullet point} in your answer.
Adjustment	Linguistic		How did US states get their names? Please respond in the writing {style of Shakespeare}.
	Situation		Change the first person to {the third person} in the given sentence.
	Daily Instructions	250 (En)	{I have a small child at home}. How can I increase my productivity while working from home?
		250 (Cn)	Write a memoir about campus life. Hmm, change it to a graduation speech.
			Write a product promotional copy. Ah, hold on, I would like you to write a user guide, with clear steps.
			Write a funny joke. Oh no, it should be a heartwarming short story.

Table 1: The statistics and examples of the SpeechInstructBench.

#### 4.1 Closed-Ended Instruction

Closed-ended instruction questions objectively evaluate SpeechLLM’s ability to follow predefined rules using IFEval (Zhou et al., 2023) metrics: Prompt-level strict-accuracy, Prompt-level loose-accuracy, Instruction-level strict-accuracy, and Instruction-level loose-accuracy. Specifically, we adapt the English IFEval (Zhou et al., 2023) dataset by selecting 432 questions suited for speech modality. We create corresponding 402 Chinese questions and modify them to accommodate spoken language-specific instructions.

All SpeechInstructBench speech data is generated using the Azure AI Speech system. To capture the complexity of real-world speech interactions, we synthesize speech data for the closed-ended instruction incorporating the following elements: 1) Accents: we implement 13 English and 8 Chinese regional accents, such as Indian and Irish variants for English, while Cantonese and Sichuan dialects for Chinese; 2) Environmental Noise: we integrate 80 distinct background sounds (e.g., markets, subway stations, background music) combined with clean speech to simulate the real-world acoustic conditions; 3) Paralinguistic Features: we vary six emotions (happy, sad, terrified, angry, unfriendly, general), multiple timbres (20 in Chinese, 25 in English), and diverse prosodic parameters (speech rate, pitch, energy); 4) Spontaneous Disfluencies: based on the disfluency definitions provided in (Ferreira and Bailey, 2004), we utilize GPT-4o with designed prompts to generate corresponding disfluent versions of instructions, incorporating six disfluency types: filled pauses, unfilled pauses, repetitions, self-corrections, false starts, and filler words. The detailed statistics and prompts are provided in Appendix B.

#### 4.2 Open-Ended Instruction

For more complex instruction-following tasks that extend beyond deterministic program verification, such as writing "in Shakespeare’s style" or "providing examples," we build up an open-ended instruction set. Each element is decomposed into binary sub-questions and could be answered by "YES" for criterion adherence or "NO" for non-compliance. We employ GPT-4o as the evaluator and the evaluation metrics for open-ended instructions are consistent with closed-ended instructions, except with the loose criteria.

For data construction, we collect raw data from previous works (Jiang et al., 2024; Qin et al., 2024) and curate 1000 instructions, along with 5,000 decomposed sub-questions, to construct the final closed-ended instruction set. The set covers five categories: 1) content (i.e., explicit restrictions on the response content); 2) format (i.e., response format requirements); 3) style (i.e., response style requirements); 4) linguistic (i.e., dictate the use of particular language structures and terms); and 5) situation (i.e., specific situation/background information added to the question).

#### 4.3 Adjustment Instruction

In practical applications, users often need to adjust or correct instructions. This scenario involves two key components: the original instruction (termed the erroneous instruction) and its revised version (the modified instruction). To meet this need, we develop an Adjustment Instruction set for the modification or correction of instructions. Using GPT-4o, we generate a set of 500 adjustment instructions and introduce two evaluation metrics: Instruction Adherence Rate (IAR) and Error Correction Rate (ECR).  $IAR = \frac{1}{n} \sum_{i=1}^n f(resp_i, inst_i)$ , where  $n$  is the total number of instructions,  $resp_i$  is the model’s response to the  $i$ -th instruction, and

SpeechLLMs	AlpacaEval	CommonEval	SD-QA		MMSU	OepnBookQA	AdvBench
	GPT↑ (S / T)	GPT↑ (S / T)	Panda↑ (S / T)	GPT↑ (S / T)	ACC↑ (S / T)	ACC↑ (S / T)	RR↑ (S / T)
BLSP	3.76 / 4.13	3.13 / 4.17	46.65 / 62.92	23.83 / 41.24	25.01 / 37.57	21.97 / 52.08	8.65 / 7.88
GLM-4-Voice	3.97 / 4.57	3.42 / 4.10	52.44 / 59.49	36.98 / 38.65	39.75 / 42.81	53.41 / 61.09	88.08 / 85.00
Mini-Omni	1.95 / 2.23	2.02 / 2.55	23.69 / 26.04	4.16 / 7.23	24.69 / 26.74	26.59 / 30.55	37.12 / 86.35
Mini-Omni2	2.32 / 2.65	2.18 / 2.86	11.03 / 13.02	7.59 / 9.76	24.27 / 27.13	26.59 / 32.09	57.50 / 92.88
Megrez	3.50 / 4.23	2.95 / 3.90	34.53 / 54.61	25.95 / 53.12	27.03 / 49.93	28.35 / 57.58	87.69 / 96.15
DIVA	3.67 / <b>4.68</b>	3.54 / <b>4.29</b>	<b>62.39 / 78.30</b>	51.72 / <b>74.50</b>	25.76 / 63.31	25.49 / 76.70	<b>98.27 / 99.23</b>
Qwen2-Audio	3.74 / 4.11	3.43 / 3.77	41.77 / 61.66	29.66 / 40.69	35.72 / 45.02	49.45 / 67.91	96.73 / 96.73
<b>InSertter</b>	<b>4.23</b> / 4.39	<b>3.63</b> / 4.05	59.13 / 61.66	<b>57.32</b> / 64.20	<b>59.27</b> / <b>64.03</b>	<b>77.14</b> / <b>83.52</b>	97.69 / 97.50

Table 2: The performance of SpeechLLMs on VoiceBench. T and S refer to the model performance with text-form and speech-form input respectively. For AdvBench, RR stands for Refusal Rate.

$inst_i$  is the  $i$ -th modified instruction. The function  $f(resp_i, inst_i)$  equals 1 when the response adheres to the modified instruction, and 0 otherwise.  $ECR = \frac{1}{m} \sum_{j=1}^m g(resp_j, inst_j)$ , where  $m$  is the total number of erroneous instructions,  $resp_j$  is the model’s response to the  $j$ -th erroneous instruction, and  $inst_j$  is the  $j$ -th erroneous instruction.  $g(resp_j, inst_j) = 1$  if the model successfully corrects the erroneous instruction, and 0 otherwise.

Notably, for both open-ended and adjustment instruction sets, we conduct human evaluations and compare them with GPT-4o judgments, demonstrating consistent agreement with over 92% accuracy across randomly selected cases. This will ensure the reliability of our evaluation results.

## 5 Experiments

### 5.1 Experimental Setup

In this study, Qwen2-Audio-7B (Chu et al., 2024) is used as the base Language Model (LLM) for all experiments, where Whisper-Large-v3 (Radford et al., 2022) encoder is used as the speech encoder to extract embeddings. The input speech segments are all resampled with a 16kHz sample rate and converted into 128-channel mel-spectrograms with a 25ms window and 10ms hop size, following a pooling layer with a stride of two.

For pre-training, we add 40% speech-text interleaved data via InSertter, combined with other speech and text data. The model is trained with a global batch size of 1024 and the sequence length is fixed at 8192 tokens. Despite the increase in data volume, the model is trained for one epoch to ensure that the training amount on the original dataset is equivalent to that of the base model, facilitating a fair comparison. For supervised fine-tuning (SFT), we utilize the same dialogue dataset for all

experiments to investigate the performance gains attributed only from InSertter pre-training. The fine-tuning process runs for one epoch. The Adam optimizer is used with a learning rate of  $1e^{-5}$ .

### 5.2 Main Results

We evaluate InSertter on VoiceBench (Chen et al., 2024b) and SpeechInstructBench against several previous state-of-the-art SpeechLLMs, including Qwen2-Audio (Chu et al., 2024), GLM-4-Voice (Zeng et al., 2024a), Mini-Omni (Xie and Wu, 2024a), Mini-Omni2 (Xie and Wu, 2024b), BLSP (Wang et al., 2024), DIVA (Held et al., 2024), and Megrez-3B-Omni (Infinigence AI, 2024).

**VoiceBench results.** We adopt multiple evaluation tasks from VoiceBench (Chen et al., 2024b) to comprehensively evaluate our model’s general speech processing capabilities. Across these diverse tasks shown in Table 2, InSertter achieves competitive performance on multiple key metrics. In particular, our model demonstrates superior performance on AlpacaEval and CommonEval with GPT-scores of 4.23 and 3.63 in speech modality, respectively, and particularly excels on OpenBookQA with 77.14% accuracy, substantially outperforming Qwen2-Audio (49.45%) and other models. Additionally, InSertter maintains strong performance on benchmarks such as MMSU and AdvBench, showing its robust and balanced capabilities across various speech-related tasks.

**SpeechInstructBench results.** In SpeechInstructBench, evaluation metrics encompass prompt-level (P) and instruction-level (I) accuracy for both closed-ended and open-ended tasks. For closed-ended questions, P and I metrics are computed by averaging loose and strict accuracies, while adjustment task is assessed using Instruction Adherence

SpeechLLMs	Closed-Ended					Open-Ended	Adjustment
	Standard (P/I. Acc) $\uparrow$	Background (P/I. Acc) $\uparrow$	Accent (P/I. Acc) $\uparrow$	Paralinguistics (P/I. Acc) $\uparrow$	Disfluency (P/I. Acc) $\uparrow$	Standard (P/I. Acc) $\uparrow$	Standard (IAR $\uparrow$ / ECR $\downarrow$ )
<i>English</i>							
BLSP	14.97 / 24.17	13.76 / 23.46	13.28 / 22.89	13.65 / 23.60	13.10 / 22.88	11.78 / 21.55	35.45 / 35.05
GLM-4-Voice	18.28 / 29.39	17.29 / 27.89	20.37 / 31.11	20.04 / 31.75	18.83 / 30.18	28.17 / 52.05	77.91 / 24.89
Mini-Omni2	7.04 / 16.52	7.26 / 16.52	7.80 / 16.91	6.16 / 15.59	7.92 / 16.30	3.23 / 6.84	12.80 / 21.20
Mini-Omni	8.14 / 16.73	9.25 / 17.74	8.37 / 16.88	8.14 / 18.02	8.23 / 17.52	1.15 / 1.70	7.31 / 18.69
Megrez	19.49 / 31.04	17.51 / 28.96	18.31 / 29.18	19.02 / 30.53	19.28 / 28.67	37.27 / 64.35	55.06 / 31.98
DIVA	27.64 / 37.26	26.32 / 36.69	26.49 / 36.26	27.97 / 37.91	19.16 / 27.89	33.64 / 61.12	58.94 / 33.73
Qwen2-Audio	19.82 / 30.18	18.17 / 28.82	18.59 / 28.81	20.70 / 31.33	15.19 / 24.67	31.40 / 58.14	48.60 / 37.45
<b>InSertter</b>	<b>39.75 / 51.35</b>	<b>37.56 / 49.87</b>	<b>37.34 / 48.24</b>	<b>35.79 / 46.85</b>	<b>36.38 / 47.28</b>	<b>40.87 / 67.33</b>	<b>80.72 / 23.28</b>
<i>Chinese</i>							
GLM-4-Voice	18.31 / 26.72	16.58 / 23.99	12.64 / 20.86	17.28 / 25.76	14.51 / 22.87	39.03 / 56.52	81.27 / 13.15
Megrez	18.31 / 27.84	17.16 / 26.32	17.26 / 26.01	18.43 / 27.92	16.58 / 25.76	31.69 / 39.27	63.34 / 17.53
DIVA	15.86 / 24.39	17.05 / 25.92	11.71 / 21.18	15.55 / 24.87	14.97 / 24.31	10.62 / 31.10	25.94 / 13.39
Qwen2-Audio	19.23 / 28.97	18.89 / 28.49	17.99 / 26.99	18.20 / 27.92	18.77 / 26.48	40.64 / 64.37	64.00 / 23.20
<b>InSertter</b>	<b>32.71 / 42.37</b>	<b>32.60 / 42.30</b>	<b>27.95 / 36.15</b>	<b>33.99 / 43.18</b>	<b>32.48 / 41.33</b>	<b>50.58 / 68.32</b>	<b>84.06 / 12.10</b>

Table 3: Performance comparison of SpeechLLMs on SpeechInstructBench. P. and I. refer to prompt-level and instruction-level accuracy, respectively. For closed-ended questions, P.&I. are calculated based on the average of both loose and strict accuracies. Adjustment tasks are evaluated using Instruction Adherence Rate (IAR) and Error Correction Rate (ECR). Notably, Mini-Omni2, Mini-Omni, and BLSP models are excluded from the Chinese benchmark comparisons due to their lack of Chinese language response capabilities.

Rate (IAR) and Error Correction Rate (ECR).

On the English benchmark (Table 3), InSertter achieves 39.75% prompt-level accuracy and 51.35% instruction-level accuracy on the standard closed-ended task, substantially surpassing the second-best performer DIVA (27.64% and 37.26%, respectively). InSertter maintains strong performance across all instruction-following tasks on SpeechInstructBench, especially on difficult scenarios (e.g., background noise), highlighting its robust capabilities. A similar pattern is observed in the Chinese benchmark (Table 3), where InSertter achieves state-of-the-art performance across all tasks. Notably, Mini-Omni2, Mini-Omni, and BLSP models are excluded from Chinese benchmark comparisons as they do not support Chinese language responses.

### 5.3 Ablation Study

#### 5.3.1 Alignment Strategy Comparison

To ensure a fair comparison of different post-training alignment algorithms (behavior alignment (Wang et al., 2024; Held et al., 2024) and representation alignment (Held et al., 2024)), we develop a baseline pre-training checkpoint derived from Qwen2-Audio. This baseline is trained without interleaved data, while maintaining the total dataset volume and pre-training model configura-

Variations	Training strategy	Closed-Ended
		(P/I. Acc) $\uparrow$
Baseline	W/o Inter. data	23.40 / 31.75
BLSP	Continuation Writing	28.20 / 39.27
DIVA	Distillation	26.81 / 37.58
<b>InSertter</b>	Sentence-Level	31.08 / 42.98
	Word-Level	<b>36.56 / 47.38</b>
<b>InSertter</b>	+ Distillation	35.56 / 46.38
	+ CW. & Distillation	38.78 / 50.96
	+ Continuation Writing	<b>39.75 / 51.35</b>

Table 4: We compare different alignment strategies, where "BLSP" refers to using the same alignment method as BLSP (Wang et al., 2024), "DIVA" denotes the DIVA (Held et al., 2024) alignment method, and "CW." represents continuation writing data. For our method, the pre-training experiments are conducted at both the word-level and sentence-level interleaved data. Speech instruction-following performance is evaluated using the closed-ended SpeechInstructBench dataset (English standard version).

tion consistent with InSertter. During the supervised fine-tuning (SFT) phase, in addition to training on the same dialogue data as InSertter, we incorporate different post-training alignment methods' training tasks and data. The mixing ratio is carefully adjusted, with experiments conducted at intervals of 10% ranging from 10% to 90%, to ensure that these algorithms achieve optimal perfor-

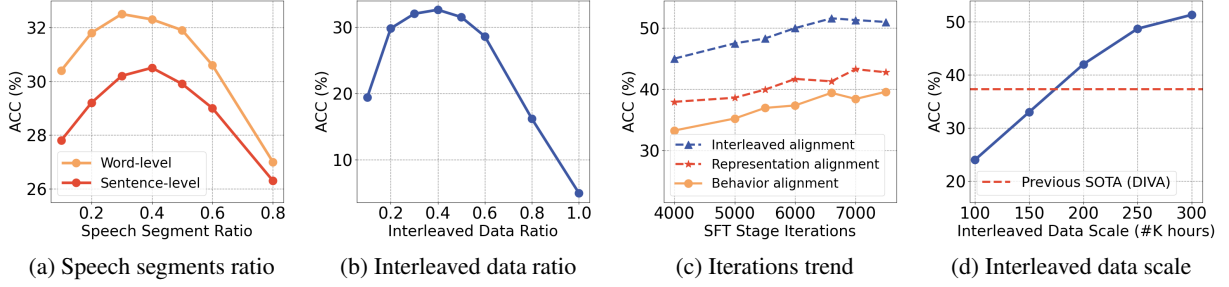


Figure 3: (a) Performance variation with different speech segment ratios. (b) Impact of interleaved data proportion on performance. (c) Different alignment strategies performance analysis with iterations during the SFT stage. (d) Performance benefits from expanding data scale. All accuracy scores refer to the closed-ended SpeechInstructBench (English standard-version) instruction-level accuracy.

mance. Unless otherwise specified, we employ the same settings for all baselines and InSertter, including training hyperparameters. For post-training datasets, we construct the same amount of data for BLSP (Wang et al., 2024) and DIVA (Held et al., 2024) experiments to match their original implementations: 8.8 million (speech, text continuation) pairs from GigaSpeech (Chen et al., 2021) and Wenetspeech (Zhang et al., 2022) datasets for BLSP, and 3.5 thousand hours of CommonVoice 17 (Ardila et al., 2020) data for DIVA.

Table 4 demonstrates InSertter’s superior performance when using word-level interleaved data. Compared to existing alignment methods and the baseline, it achieves significant improvements of 36.56% in prompt-level accuracy and 47.38% in instruction-level accuracy. The word-level InSertter approach proves more effective compared with sentence-level as it aligns better with the text continuation objective, providing finer granularity for speech-text alignment. Based on our ablation studies (Fig. 3(a)), we determined the optimal speech segment ratios to be 30% for word-level and 40% for sentence-level interleaved data.

Moreover, model performance gains are further enhanced when combining interleaved pre-training strategy with existing post-training alignment methods. Specifically, integrating word-level interleaved pre-training method with continuation writing post-training strategy yields additional improvements of 39.56% (prompt-level accuracy) and 51.35% (instruction-level accuracy). We use this version as the final competitive model for the main results section (Sec. 5.2).

### 5.3.2 Hyperparameters and Data Scaling Analysis

To rigorously evaluate the effectiveness of our InSertter approach, we study the effects of data scal-

ing and InSertter related hyperparameters. First, as shown in Fig. 3(a), we conduct experiments on the pre-training stage to determine the optimal speech segment ratio for interleaved data. The results reveal that model performance peaks when speech segment ratio falls between 20% and 40%, with the highest accuracy achieved around 30% for word-level and 40% for sentence-level interleaved data. As presented in Fig. 3(b), varying the interleaved data ratio among pre-training data greatly influences performance, with an optimal ratio nearing 0.4 for achieving the highest accuracy. Fig. 3(c) presents the impact of iterations during the supervised fine-tuning (SFT) stage, indicating that interleaved alignment outperforms representation and behavior alignments consistently over increasing iterations, peaking between 6000 and 7000 iterations. So we set SFT with 7000 iterations for all experiments. Finally, Fig. 3(d) illustrates the relationship between interleaved data scale and instruction-following performance, confirming that larger interleaved data scales lead to substantial performance improvements, ultimately surpassing the previous state-of-the-art accuracy (as indicated by the red dashed line for DIVA (Held et al., 2024)) when data scale reaches around 300K hours. This proves that our proposed method can scale effectively and yield more performance benefits by continuously expanding the interleaved data.

## 6 Conclusion

In this paper, we introduced a novel and easily scalable pre-training framework termed InSertter, alongside a comprehensive speech instruction-following benchmark. Our proposed InSertter demonstrates state-of-the-art performance across various instruction-following tasks while maintaining competitive results on diverse speech-language



tasks. Future work would explore more efficient training techniques and investigate the potential of extending our approach to more diverse languages and speech scenarios.

## 7 Limitation

Despite the promising results obtained in our method, there are still several areas for improvement. Firstly, the training process consists of only English and Chinese related datasets. The generalization to other languages is yet to be verified. Secondly, for evaluation, some tasks of SpeechInstructBench (open-ended instruction, adjustment instruction) heavily rely on GPT-4o API for scoring. However, the availability and accessibility of the GPT-4o API are external factors beyond our control. Addressing these limitations in future work would provide a more comprehensive validation of our approach.

## 8 Ethical Consideration

We affirm that we contribute to society, avoid harm, and are honest and trustworthy. We respect previous work and appropriately cite the methods and datasets we are using. All the data we use is subject to strict verification to ensure it contains no harmful or sensitive content.

## References

- Pavel Andreev, Aibek Alanov, Oleg Ivanov, and Vetrov Dmitry. 2023. [Hifi++: A unified framework for bandwidth extension and speech enhancement](#). In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 1–5. IEEE.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France. European Language Resources Association.
- Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audioldm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.
- Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Yujun Wang, Zhao You, and Zhiyong Yan. 2021. Giga-speech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Proc. Interspeech 2021*.
- Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, Dingdong Wang, Kun Xiang, Haoyuan Li, Haoli Bai, Jianhua Han, Xiaohui Li, Weike Jin, Nian Xie, Yu Zhang, James T. Kwok, Hengshuang Zhao, Xiaodan Liang, Dit-Yan Yeung, Xiao Chen, Zhenguo Li, Wei Zhang, Qun Liu, Jun Yao, Lanqing Hong, Lu Hou, and Hang Xu. 2024a. [Emova: Empowering language models to see, hear and speak with vivid emotions](#). *Preprint*, arXiv:2409.18042.
- Qian Chen, Yunfei Chu, Zhifu Gao, Zerui Li, Kai Hu, Xiaohuan Zhou, Jin Xu, Ziyang Ma, Wen Wang, Siqi Zheng, et al. 2023. Lauragpt: Listen, attend, understand, and regenerate audio with gpt. *arXiv preprint arXiv:2310.04673*.
- Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. 2024b. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.
- Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.
- Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. [Moshi: a speech-text foundation model for real-time dialogue](#). Technical report.
- Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024. [Cosyvoice 2: Scalable streaming speech synthesis with large language models](#). *Preprint*, arXiv:2412.10117.
- Fernanda Ferreira and Karl GD Bailey. 2004. Disfluencies and human language comprehension. *Trends in cognitive sciences*, 8(5):231–237.
- Kuofeng Gao, Shu-Tao Xia, Ke Xu, Philip Torr, and Jindong Gu. 2024. [Benchmarking open-ended audio dialogue understanding for large audio-language models](#). *Preprint*, arXiv:2412.05167.

- Yuan Gong, Alexander H Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. 2023a. Joint audio and speech understanding. In *ASRU*, pages 1–8. IEEE.
- Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and Glass James. 2023b. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.
- Will Held, Ella Li, Michael Ryan, Weiyan Shi, Yanzhe Zhang, and Diyi Yang. 2024. Distilling an end-to-end voice assistant from speech recognition data.
- Infinigence AI. 2024. Megrez-3b-omni. <https://github.com/infinigence/Infini-Megrez-Omni>.
- Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, et al. 2024. Wavchat: A survey of spoken dialogue models. *arXiv preprint arXiv:2411.13577*.
- Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024. **Follow-Bench: A multi-level fine-grained constraints following benchmark for large language models**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688, Bangkok, Thailand. Association for Computational Linguistics.
- Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Du, et al. 2024. An embarrassingly simple approach for llm with strong asr capacity. *arXiv preprint arXiv:2402.08846*.
- Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. 2024. **Spirit-lm: Interleaved spoken and written language model**. *Preprint*, arXiv:2402.05755.
- Beijing Academy of Artificial Intelligence (BAAI). 2024. Infinity instruct. *arXiv preprint arXiv:2406.XXXX*.
- Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. **Infobench: Evaluating instruction following ability in large language models**.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. **Robust speech recognition via large-scale weak supervision**. *arXiv preprint*.
- Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.
- Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2024. **Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents**. *Preprint*, arXiv:2305.13040.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. **SALMONN: Towards generic hearing abilities for large language models**. In *The Twelfth International Conference on Learning Representations*.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca).
- Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. 2024. **Beyond turn-based interfaces: Synchronous llms as full-duplex dialogue agents**. *Preprint*, arXiv:2409.15594.
- Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. 2024. **Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing**. *Preprint*, arXiv:2309.00916.
- Zhifei Xie and Changqiao Wu. 2024a. **Mini-omni: Language models can hear, talk while thinking in streaming**. *Preprint*, arXiv:2408.16725.
- Zhifei Xie and Changqiao Wu. 2024b. **Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities**. *ArXiv*, abs/2410.11190.
- Hongfei Xue, Yuhao Liang, Bingshen Mu, Shiliang Zhang, Mengzhe Chen, Qian Chen, and Lei Xie. 2024. **E-chat: Emotion-sensitive spoken dialogue system with large language models**. *Preprint*, arXiv:2401.00475.
- Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. 2024. **Air-bench: Benchmarking large audio-language models via generative comprehension**. *arXiv preprint arXiv:2402.07729*.
- Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024a. **Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot**. *Preprint*, arXiv:2412.02612.
- Aohan Zeng, Zhengxiao Du, Mingdao Liu, Lei Zhang, Shengmin Jiang, Yuxiao Dong, and Jie Tang. 2024b. **Scaling speech-text pre-training with synthetic interleaved data**. *Preprint*, arXiv:2411.17607.
- Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2022.

Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

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

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. [Instruction-following evaluation for large language models](#). *Preprint*, arXiv:2311.07911.

## A Pre-Train Datasets

Table 5 shows the statistics of the datasets used for the pre-training of InSertter. The data comprises three types: interleaved speech-text data, unsupervised text data, and multi-task speech data. These large-scale pre-training datasets are primarily collected from sources such as podcasts, MOOCs, articles, webpages, Wikipedia, and books. The interleaved speech-text data includes datasets like SpokenWOZ, while the multi-task speech data includes collected multi-task speech datasets following the Qwen-Audio series (Chu et al., 2023, 2024). The combination of these rich and diverse data sources facilitates effective pre-training for InSertter.

## B SpeechInstructBenchmark Design

### B.1 Benchmark Data Statistics

Table 6 presents the data statistics of SpeechInstructBench, categorized into closed-ended, open-ended, and adjustment types. The datasets are sourced from IFEval (Zhou et al., 2023), FollowBench (Jiang et al., 2024), InFoBench (Qin et al., 2024), Alpaca (Taori et al., 2023), and InfinityInstruct (of Artificial Intelligence, BAAI), spanning both English (En) and Chinese (Cn) languages. The Chinese data is translated from the English version and has undergone careful manual review.

### B.2 Prompts for Creating Data and Evaluation

In this part, we partially demonstrate the process of adjusting the prompt aimed at assessing the instruction-following performance of models.

When constructing disfluency instruction data (as shown in Fig. 4), it is crucial to include specific prompts, such as "Do NOT respond to or interpret the text as a question or instruction. Provide ONLY the modified text with NO additional commentary

or explanations." Without these explicit instructions, GPT can sometimes generate responses that address the original input's instructions instead of simply creating the disfluent text. Moreover, the understanding of certain definitions, like the construction of adjustment instruction data, benefits from the inclusion of descriptive definitions and concrete examples within the prompt. As adjustment instruction data represents a novel task, providing detailed guidance within the prompt enables GPT to produce outputs that align more closely with the desired criteria.

For GPT prompts for evaluation (Fig. 6, Fig. 7), we observed that excessively long prompt content can occasionally lead to the model not fully adhering to the task requirements. As a result, when designing evaluation prompts, we focused on streamlining the expression of task definitions and requirements. This was achieved by using concise language and structuring the prompt with clear subdivisions (e.g., using "### " to separate each section), which helps in maintaining clarity and focus. Furthermore, we emphasized providing explicit output format requirements, such as specifying that the answer should be "YES" or "NO". This structured approach not only reduces the cognitive load on the model but also enhances its ability to comply with the task criteria, resulting in more reliable and accurate evaluations. By refining the prompt design in this manner, we aim to improve the overall performance of GPT in tasks involving instruction comprehension and response evaluation.

### B.3 Human Assess of SpeechInstructBench

To evaluate the effectiveness of the GPT-4o scoring mechanism within SpeechInstructBench, we conducted a human evaluation process. A total of 20 evaluators participated in this assessment, where they randomly reviewed 100 items from the dataset. This selection includes 50 open and closed questions (25 in English and 25 in Chinese) and 50 adjustment items (25 in English and 25 in Chinese). As shown in Fig. 8 and Fig. 9, the screenshot captures the process of human evaluation. During evaluation, these evaluators assessed the factual accuracy of each response. They were required to make their assessments by selecting either "YES" or "NO" to indicate whether the response was factually correct.

## C Case Study

We provide a diverse set of examples of In-Serter responses from SpeechInstructBench and VoiceBench (AlpacaEval, CommonEval, SD-QA, MMSU, OpenBookQA, AdvBench), as illustrated from Fig. 10 to Fig. 21.



Types	Dataset-Source	Speech Hours	Text Tokens
Interleaved speech-text data	Podcast	301540	610 B
	MOOC		
	Articles		
	SpokenWOZ (Si et al., 2024)		
Unsupervised text data	Webpages	-	450 B
	Wikipedia		
	Articles		
	Books		
Multi-task speech data	Collected multi-task speech data	226000	445 B

Table 5: The statistics of the pre-training datasets.

Prompt Template (Generating disfluent version closed-ended instruction data)
<p>### Definitions of Speech Disfluencies:</p> <ol style="list-style-type: none"> <li>1. Filled Pauses: <ul style="list-style-type: none"> <li>- Definition: Non-lexical sounds such as "uh," "um," "er," or "ah" that are inserted into speech to fill pauses, typically occurring when the speaker is uncertain or thinking.</li> </ul> </li> <li>2. Unfilled Pauses: <ul style="list-style-type: none"> <li>- Definition: Silent intervals in speech, generally occurring when the speaker is thinking or organizing their thoughts.</li> </ul> </li> <li>3. Repetitions: <ul style="list-style-type: none"> <li>- Definition: The speaker repeats a word or phrase, often for emphasis, to confirm an upcoming expression, or due to a break in thought.</li> </ul> </li> <li>4. Self-corrections: <ul style="list-style-type: none"> <li>- Definition: Occur when the speaker becomes aware of an error, inaccuracy, or inappropriate language and makes a deliberate correction or adjustment.</li> </ul> </li> <li>5. False Starts: <ul style="list-style-type: none"> <li>- Definition: The speaker begins a sentence or phrase, then changes their approach or restructures it, leaving the initial segment incomplete or interrupted.</li> </ul> </li> <li>6. Filler Words: <ul style="list-style-type: none"> <li>- Definition: Words or phrases used in daily conversation to fill gaps, reorganize thoughts, or maintain the flow of dialogue. They usually have no actual semantic meaning but help maintain the flow and tone of conversation.</li> </ul> </li> </ol> <p>### Instruction:</p> <p>Reconstruct the following text by incorporating the specified disfluency features. Do not respond or interpret the content as a standalone question or directive.</p> <ul style="list-style-type: none"> <li>- Filled pauses should occur: {number_filled_pause} times.</li> <li>- Unfilled pauses (indicated by " ... ") should occur: {number_unfilled_pause} times.</li> <li>- Repetitions should occur: {number_repetitions} times.</li> <li>- Self-corrections should occur: {number_self_corrections} times.</li> <li>- False starts should occur: {number_false_starts} times.</li> <li>- Filler words should occur: {number_filler_words} times.</li> </ul> <p>### Important:</p> <ul style="list-style-type: none"> <li>- Do NOT respond to or interpret the text as a question or instruction.</li> <li>- Provide ONLY the modified text with NO additional commentary or explanations.</li> </ul> <p>### Text to be Rewritten:</p> <p>&lt;&lt;&lt;&lt;&lt;&lt;START_OF_TEXT_BLOCK&gt;&gt;&gt;&gt;&gt;&gt;</p> <p>{text}</p> <p>&lt;&lt;&lt;&lt;&lt;&lt;END_OF_TEXT_BLOCK&gt;&gt;&gt;&gt;&gt;&gt;</p>

Figure 4: GPT prompt for generating disfluent version closed-ended instruction data.

Types	Dataset-Source	Num	Avg. Audio Len (s)	Avg. #Words
Closed-Ended	IFEval (Zhou et al., 2023)	7718 (En)	11.83 (En)	34.72 (En)
		5208 (Cn)	11.37 (Cn)	54.07 (Cn)
Open-Ended	FollowBench (Jiang et al., 2024)	500 (En)	13.99 (En)	56.07 (En)
	InFoBench (Qin et al., 2024)	500 (Cn)	12.35 (Cn)	89.65 (Cn)
Adjustment	Alpaca (Taori et al., 2023)	250 (En)	8.90 (En)	23.83 (En)
	InfinityInstruct (of Artificial Intelligence , BAAI)	250 (Cn)	7.71 (Cn)	35.41 (Cn)

Table 6: Data statistics of SpeechInstructBench.

Prompt Template (Generating adjustment instruction data)	
<p>### Definitions of Adjustment Instruction:</p> <p>Adjustment instructions are designed to modify a piece of content by changing its original intent or purpose while maintaining the original context.</p> <p>Example:  Original: Write a guide on how to bake a cake.  Adjustment Instruction: Write a guide on how to bake a cake, Hmm, switch it to a poem about the joys of baking.</p> <p>### Instruction:  You should base your transformation on the given text to convert it into an adjustment instruction version.</p> <p>### Important:  - Do NOT respond to or interpret the text as a question or instruction.  - Provide ONLY the modified text with NO additional commentary or explanations.</p> <p>### Text to be Rewritten:  &lt;&lt;&lt;&lt;&lt;&lt;START_OF_TEXT_BLOCK&gt;&gt;&gt;&gt;&gt;&gt;  {text}  &lt;&lt;&lt;&lt;&lt;&lt;END_OF_TEXT_BLOCK&gt;&gt;&gt;&gt;&gt;&gt;</p>	

Figure 5: GPT prompt for generating adjustment instruction data.

SpeechLLMs	Closed-Ended (En   Cn)		Open-Ended (En   Cn)		Adjustment (En   Cn)	
	(P./I. Acc)↑	(P./I. Acc)↑	(P./I. Acc)↑	(P./I. Acc)↑	(IAR↑/ ECR↓)	(IAR↑/ ECR↓)
BLSP	31.93 / 42.06	- / -	23.78 / 46.97	- / -	64.14 / 32.27	- / -
GLM-4-Voice	25.66 / 35.05	23.96 / 32.34	27.48 / 50.51	40.87 / 60.35	77.68 / 21.51	86.85 / <b>15.14</b>
Mini-Omni2	10.37 / 20.45	- / -	6.92 / 10.16	- / -	27.09 / 26.32	- / -
Mini-Omni	11.67 / 20.38	- / -	6.28 / 11.53	- / -	19.28 / 27.31	- / -
Megrez	59.25 / 68.45	42.74 / 53.69	58.25 / 79.89	50.76 / 75.34	88.92 / <b>19.49</b>	82.38 / 17.74
DIVA	<b>66.07 / 74.32</b>	<b>48.50 / 58.66</b>	<b>60.50 / 87.80</b>	<b>58.42 / 84.11</b>	<b>87.50 / 29.44</b>	<b>91.53 / 19.35</b>
Qwen2-Audio	28.63 / 38.62	33.29 / 43.25	49.79 / 75.54	47.34 / 72.26	76.70 / 25.30	84.0 / 22.40
InSertter	42.62 / 52.00	41.02 / 51.83	49.19 / 75.81	52.19 / 68.5	81.13 / 24.10	86.06 / 20.31

Table 7: Performance comparison of SpeechLLMs on SpeechInstructBench with text-form input. The P. and I. refer to prompt-level and instruction-level accuracy, respectively. For closed-ended questions, P.&I. are calculated based on the average of both loose and strict accuracies. Adjustment tasks are evaluated using Instruction Adherence Rate (IAR) and Error Correction Rate (ECR). Notably, Mini-Omni2, Mini-Omni, and BLSP models are excluded from the Chinese benchmark comparisons due to their lack of Chinese language response capabilities.

### Prompt Template (Evaluating open-ended instruction response)

#### ### Task Definition:

You are given a user instruction, and related AI assistant's answer. Your task is to judge if the AI assistant's answer correctly follows the given sub-questions. You should answer the ensuing sub-questions with either a YES or NO choice.

Your selection should be based on your judgment as well as the following rules:

- YES: Select 'YES' if the generated text entirely fulfills the condition specified in the question. However, note that even minor inaccuracies exclude the text from receiving a 'YES' rating. As an illustration, consider a question that asks, \"Does each sentence in the generated text use a second person?\" If even one sentence does not use the second person, the answer should NOT be 'YES'. To qualify for a 'YES' rating, the generated text must be entirely accurate and relevant to the question.

- NO: Select 'NO' if the generated text fails to meet the question's requirements or provides no information that could be utilized to answer the question. For instance, if the question asks, \"Is the second sentence in the generated text a compound sentence?\" and the generated text only has one sentence. It offers no relevant information to answer the question. Consequently, the answer should be 'NO'.

#### ### Original User Instruction:

{prompt}

#### ### AI Assistant's Answer:

{response}

#### ### Questions for Evaluation:

{decompose\_question}

#### ### Note:

You should only return a list consisting of 'YES' or 'NO' for each question. For example, if there are 3 questions, your list should contain 3 elements, each corresponding to the judgment for the respective question.

Figure 6: GPT prompt for evaluating open-ended instruction task.

Accent	InSertter (P/I. Acc)↑	Qwen2-Audio (P/I. Acc)↑	BLSP (P/I. Acc)↑	DIVA (P/I. Acc)↑	GLM-4-Voice (P/I. Acc)↑	Mini-Omni2 (P/I. Acc)↑	Mini-Omni (P/I. Acc)↑	Megrez (P/I. Acc)↑
en-AU	<b>39.20 / 49.35</b>	19.38 / 29.97	13.76 / 23.53	27.53 / 37.33	20.26 / 31.04	7.26 / 16.30	9.69 / 17.16	18.17 / 29.32
en-CA	<b>37.11 / 48.35</b>	17.95 / 28.39	12.88 / 24.10	26.65 / 36.19	21.80 / 32.11	7.04 / 16.02	9.03 / 17.73	17.40 / 29.54
en-GB	<b>38.19 / 49.77</b>	19.60 / 29.61	14.42 / 23.74	25.77 / 35.90	18.83 / 30.11	7.70 / 17.02	8.59 / 16.45	19.38 / 28.68
en-HK	<b>39.75 / 49.71</b>	19.68 / 29.81	14.86 / 25.32	26.10 / 35.69	18.72 / 29.82	6.16 / 16.02	8.59 / 17.73	19.50 / 31.69
en-IE	<b>36.67 / 48.21</b>	18.50 / 29.61	14.42 / 23.74	26.43 / 36.19	21.03 / 30.82	7.48 / 17.16	7.70 / 17.16	19.17 / 28.77
en-IN	<b>38.10 / 48.28</b>	19.71 / 28.82	13.76 / 23.39	27.53 / 36.83	16.40 / 26.60	7.70 / 17.02	9.03 / 16.30	17.19 / 30.13
en-KE	<b>37.22 / 48.42</b>	19.27 / 27.53	14.53 / 23.17	25.66 / 35.90	16.03 / 26.00	9.03 / 16.88	8.59 / 16.73	18.40 / 28.19
en-NG	<b>34.36 / 45.42</b>	17.62 / 28.32	13.54 / 21.38	25.35 / 34.56	19.02 / 28.66	9.69 / 18.16	8.59 / 16.02	17.96 / 29.12
en-NZ	<b>39.20 / 50.35</b>	17.47 / 27.87	13.10 / 21.67	25.99 / 36.33	17.78 / 27.85	7.48 / 16.30	9.03 / 18.02	20.61 / 31.35
en-PH	<b>39.75 / 50.78</b>	20.15 / 29.39	12.99 / 23.24	24.66 / 35.05	19.27 / 29.18	7.26 / 16.88	8.37 / 18.16	16.87 / 28.06
en-SG	<b>36.89 / 46.85</b>	14.97 / 27.53	13.10 / 21.67	24.98 / 33.97	16.96 / 26.82	7.48 / 17.31	6.61 / 17.31	16.88 / 26.43
en-TZ	<b>35.02 / 46.49</b>	18.28 / 27.75	10.35 / 21.38	26.62 / 35.96	16.51 / 25.46	7.70 / 17.88	7.92 / 17.73	19.86 / 31.02
en-ZA	<b>36.34 / 48.06</b>	18.50 / 27.46	11.34 / 21.10	25.62 / 34.80	19.93 / 30.25	8.14 / 16.45	8.37 / 16.59	18.31 / 28.92

Table 8: Closed-Ended Instruction - English Accents Benchmark. This table evaluates model performance across various English accents including Australian (en-AU), Canadian (en-CA), British (en-GB), Hong Kong (en-HK), Irish (en-IE), Indian (en-IN), Kenyan (en-KE), Nigerian (en-NG), New Zealand (en-NZ), Filipino (en-PH), Singaporean (en-SG), Tanzanian (en-TZ), and South African (en-ZA).

### Prompt Template (Evaluating adjustment instruction response)

#### ### Task Definition:

You are given a user instruction (with adjustment style in daily spoken language tone), and related AI assistant's answer. Your task is to evaluate if the AI assistant follows the adjusted instruction.

#### ### Evaluation Guidelines:

##### 1. follow\_after\_adjust:

Check if the model correctly followed the corrected instruction after the adjustment.

- YES: The model follows the adjusted instruction.
- NO: The model fails to follow the adjusted instruction.

##### 2. follow\_before\_adjust:

Check if the model initially followed the original instruction before the adjustment.

- YES: The model followed the original instruction.
- NO: The model did not follow the original instruction.

#### Example:

Input instruction: Write a funny joke. Oh, change it to a heartwarming short story.

If the response content is about a funny joke instead of a heartwarming story, 'follow\_before\_adjust' should be 'YES', and 'follow\_after\_adjust' should be 'NO'. And if the answer is a heartwarming story but is very long, 'follow\_after\_adjust' should also be 'NO' and 'follow\_before\_adjust' is 'NO'.

#### ### Original User Instruction:

{text}

#### ### AI Assistant's Answer:

{response}

#### ### Output Format:

Return your evaluation results in the following JSON structure:

```
{
  "follow_after_adjust": "YES" or "NO",
  "follow_before_adjust": "YES" or "NO"
}
```

Figure 7: GPT prompt for evaluating adjustment instruction task.

Accent	InSertor (P./I. Acc)↑	Qwen2-Audio (P./I. Acc)↑	DIVA (P./I. Acc)↑	GLM-4-Voice (P./I. Acc)↑	Mini-Megrez (P./I. Acc)↑
cn-Wuu	<b>16.93 / 24.95</b>	13.59 / 22.39	9.21 / 18.13	10.02 / 18.21	15.32 / 23.27
cn-Yue	<b>24.30 / 32.82</b>	17.28 / 25.28	11.98 / 22.55	11.52 / 19.98	17.14 / 24.73
cn-Guangxi	<b>32.83 / 42.05</b>	19.35 / 28.49	13.36 / 23.03	14.86 / 23.19	19.79 / 27.69
cn-Henan	<b>22.92 / 30.41</b>	16.93 / 27.04	10.36 / 19.26	9.67 / 16.69	16.35 / 25.86
cn-Liaoning	<b>36.05 / 45.90</b>	21.38 / 30.62	13.94 / 23.75	17.16 / 26.24	17.28 / 26.18
cn-Shanxi	<b>18.66 / 27.12</b>	16.47 / 26.40	10.36 / 19.58	10.36 / 17.73	17.44 / 27.09
cn-Shandong	<b>27.99 / 36.27</b>	19.12 / 27.04	10.25 / 19.26	9.56 / 16.93	16.01 / 24.73
cn-Sichuan	<b>23.27 / 35.15</b>	17.39 / 26.32	10.94 / 19.90	9.67 / 17.90	17.77 / 26.52

Table 9: Closed-Ended Instruction - Chinese Accents Benchmark. This table evaluates model performance based on varying Chinese dialects: Jiang-Zhe speech (cn-Wuu), Cantonese (cn-Yue), Guangxi Mandarin (cn-Guangxi), Heman Mandarin (cn-Henan), Liaoning Mandarin (cn-Liaoning), Shanxi Mandarin (cn-Shanxi), Shandong Mandarin (cn-Shandong), Sichuan Mandarin (cn-Sichuan).



▼ Guidelines

**Instruction:**  
You are given a user instruction, and related AI assistant's answer. Your tasks is to judge if the AI assistant's answer correctly follows the given sub-questions. You should answer the ensuing sub-questions with either a YES or NO choice.

- YES: Select 'YES' if the generated text entirely fulfills the condition specified in the question. However, note that even minor inaccuracies exclude the text from receiving a 'YES' rating. As an illustration, consider a question that asks. "Does each sentence in the generated text use a second person? " If even one sentence does not use the second person, the answer should NOT be 'YES'. To qualify for a 'YES' rating, the generated text must be entirely accurate and relevant to the question.

- NO: Select 'NO' if the generated text fails to meet the question's requirements or provides no information that could be utilized to answer the question. For instance, if the question asks. "Is the second sentence in the generated text a compound sentence?" and the generated text only has one sentence. It offers no relevant information to answer the question. Consequently, the answer should be 'NO'.

<p><b>Original Instruction:</b></p> <p>Change the first person to the third person in the given sentence. The meaning should be kept, but you can paraphrase it or expand it in order to have a better pose. The given information is as follows: We were recently able to increase the amount of stock we hold with the same supplier thereby reducing our risk.</p> <p><b>Response:</b></p> <p>They were recently able to increase the amount of stock they hold with the same supplier, thereby reducing their risk.</p>	<p><b>Select an option:</b></p> <p>Is the generated text expressed in third person?:</p> <p><input type="radio"/> Yes</p> <p><input type="radio"/> No</p> <p>Does the generated text have a better pose than the given input?</p> <p><input type="radio"/> Yes</p> <p><input type="radio"/> No</p> <p>Does the generated text convey the same meaning as the original sentence in the given input?</p> <p><input type="radio"/> Yes</p> <p><input type="radio"/> No</p>
---	---

Figure 8: Screenshot of human evaluation for the open-ended instruction task.

▼ Guidelines

**Instruction:**  
Your task is to evaluate whether the given response adheres to both the adjusted instruction ('follow\_after\_adjust') and the original instruction prior to the adjustments ('follow\_before\_adjust').

**Example:**  
Input instruction: Write a funny joke. Oh, change it to a heartwarming short story.  
If the response content is about a funny joke instead of a hearwarming story, the 'follow\_before\_adjust' should be 'YES', and 'follow\_after\_adjust' should be 'NO'. And if the answer is a heartmaring story but is very long, 'follow\_after\_adjust' should also be 'NO' and 'follow\_before\_adjust' is 'NO'.

<p><b>Original Instruction:</b></p> <p>Generate a poem about summer. Hold on, I need to adjust my request, change it to a poem about autumn.</p> <p><b>Response:</b></p> <p>In autumn's gentle, crisp embrace, Leaves descend with fiery grace. Golden fields and harvest's yield, Nature's beauty is revealed. Amber skies and whispered breeze, Painted leaves on ancient trees. In autumn's glow, we find peace, As the vibrant colors cease.</p>	<p><b>Select an option:</b></p> <p>If the response correctly adjusted instruction:</p> <p><input type="radio"/> Yes</p> <p><input type="radio"/> No</p> <p>If the response follow the original instruction prior to the adjustments:</p> <p><input type="radio"/> Yes</p> <p><input type="radio"/> No</p>
--	---

Figure 9: Screenshot of human evaluation for the adjustment instruction task.

**Instruction:**

Compose a poem that has the word \"land\" and \"river\". It should be about nature and love. Also, the word \"forest\" should appear at least three times, and be written in English, with all letters lowercased.

**Response:**

in the land of love, where rivers flow  
a forest thrives, its trees so tall  
where hearts entwine, and souls do  
grow  
a place where love can truly call  
the river runs, its waters clear  
a symbol of the love we hold so  
dear  
it flows and flows, without a fear  
a constant reminder of our love's dear  
the forest stands, its  
branches wide  
a place where love can truly hide  
where secrets are kept, and hearts abide  
a place where  
love can truly abide  
in the land of love, where rivers flow  
the forest stands, its beauty to show  
where love  
can grow, and hearts can glow  
a place where love can truly grow.

**Instruction:**

Students are travelling to the University of California, Irvine for three days. Create a hilarious itinerary for them. Do not use the word 'university'. Your entire response should have exactly four paragraphs. Separate paragraphs with the markdown divider: three asterisks.

**Response:**

Day 1: The students will start their journey with a visit to the famous Disneyland Park. They will spend the day exploring the magical world of Disney, from the thrilling rides to the enchanting shows. In the evening, they will head to the nearby Universal CityWalk for some shopping and dining at the many restaurants and bars.

\*\*\*

Day 2: The students will take a day trip to the beautiful city of San Diego. They will visit the iconic San Diego Zoo, home to over 3,500 animals from around the world. Afterward, they will head to the USS Midway Museum, where they can learn about the history of the United States Navy. In the evening, they will enjoy a seafood dinner at one of the many restaurants in the Gaslamp Quarter.

\*\*\*

Day 3: The students will spend their last day in California at the famous Hollywood Walk of Fame. They will take a guided tour of the famous landmarks, including the Hollywood Sign, the TCL Chinese Theatre, and the Dolby Theatre. In the evening, they will head to the Griffith Observatory for a breathtaking view of the city and the stars.

\*\*\*

Day 4: The students will return home with memories of a trip filled with laughter, adventure, and fun. They will have experienced the best of California, from the theme parks to the beaches, and everything in between.

Figure 10: SpeechInstructBench Task: Closed-ended instruction following (English version)

**Instruction:**

我是一个十二年级的学生，我需要一些关于大学申请的帮助，你能给我一些建议吗？你的回答最后应该是“按照上述步骤你不会失败”。这句话后面不应该再有其他的话。

**Response:**

当然可以！首先，了解你感兴趣的大学和专业是非常重要的。你可以通过大学的官方网站、排名、课程设置和校友访谈来了解更多信息。其次，准备你的标准化考试成绩，如SAT或ACT，这是许多大学录取过程中的关键因素。此外，积极参与课外活动和社区服务，这不仅能展示你的领导能力和团队合作精神，还能帮助你在申请中脱颖而出。最后，撰写一份引人注目的个人陈述，这是你向大学展示自己独特性和个人故事的机会。记住，保持积极的态度，相信自己，你一定会成功的！按照上述步骤你不会失败。

**Instruction:**

纽约展望公园的历史是什么？请将您的整个答案用JSON格式包含起来。

**Response:**

```
```json\n{\n  \"park_name\": \"纽约展望公园\", \n  \"history\": \"纽约展望公园（Prospect Park）位于美国纽约市布鲁克林区，是纽约市最大的城市公园之一。公园的历史可以追溯到19世纪初，当时该地区是一片未开发的森林和农田。1867年，布鲁克林市议会决定将这片土地改造成公园，以供市民休闲和娱乐。公园的设计者是弗雷德里克·奥姆斯特德，他是纽约中央公园的设计者之一。奥姆斯特德的设计理念是将公园与周围的自然环境融为一体，创造一个既美丽又实用的空间。展望公园的建设历时数年，最终在1872年对外开放。自那时以来，公园一直是布鲁克林市民和游客的热门目的地，提供了许多户外活动和娱乐设施，包括自行车道、步道、湖泊、动物园、博物馆和剧院等。展望公园的历史和文化价值使其成为纽约市的重要地标之一。\"\n}\n```
```

**Instruction:**

写一首关于在波托马克河被一块鸡肉噎住的歌。请将标题用书名号括起。

**Response:**

《波托马克河上的鸡肉之歌》

Figure 11: SpeechInstructBench Task: Closed-ended instruction following (Chinese version)

**Instruction:**

A confirmation email should be written appropriately for the situation. \n\nThe given information is as follows: A meeting has been scheduled, and the sender expects the other to review the slides.

**Decomposed Questions:**

1. Is the generated text a confirmation email? [Yes]
2. Is the generated email relevant to a scheduled meeting? [Yes]
3. Does the generated email mention that the receiver is expected to review the slides? [Yes]

**Response:**

Subject: Confirmation of Meeting and Request for Slide Review\n\nDear [Recipient's Name],\n\nI hope this email finds you well. I am writing to confirm our upcoming meeting scheduled for [Date and Time]. I am looking forward to discussing [Meeting Topic] with you.\n\nAs part of our preparation for the meeting, I would like to request your review of the slides that I have prepared. The slides cover [Brief Description of the Slide Content] and I believe they will be helpful in our discussion.\n\nPlease let me know if you have any questions or concerns regarding the slides. I am available to make any necessary revisions or provide additional information.\n\nThank you for your time and effort in preparing for this meeting. I am confident that our discussion will be productive and beneficial for both of us.\n\nBest regards,\n\n[Your Name]", "prompt": "A confirmation email should be written appropriately for the situation. \n\nThe given information is as follows: A meeting has been scheduled, and the sender expects the other to review the slides.

**Instruction:**

Design a programming problem related to the subject that has been given to you. Use some examples and constraints to improve your question. The example should be helpful for explanation. \n\nThe given information is as follows: Dynamic Programming.

**Decomposed Questions:**

1. Does the generated text contain a programming problem? [Yes]
2. Does the generated text contain some examples of a programming problem? [Yes]
3. Does the generated text contain some constraints of a programming problem? [Yes]
4. Is the generated programming problem relevant to the subject in the given input? [Yes]
5. Are the generated examples helpful in explaining the generated programming problem? [Yes]
6. Are the generated constraints reasonable for the generated programming problem? " [Yes]

**Response:**

Problem Statement:\n\nYou are given a sequence of integers, and you need to find the maximum sum of a subsequence such that no two elements in the subsequence are adjacent in the original sequence. For example, if the sequence is [2, 4, 6, 2, 5], the maximum sum of a subsequence with no adjacent elements is 12 (2 + 6 + 4). \n\nConstraints:\n\n- The length of the sequence is between 1 and 1000. \n- Each element in the sequence is between -1000 and 1000. \n\nExamples:\n\n- Input: [2, 4, 6, 2, 5]\n Output: 12\n- Input: [5, 1, 1, 5]\n Output: 10\n- Input: [-2, -1, -3]\n Output: 0\n\nYou need to implement a function `max\_sum\_subsequence` that takes a list of integers as input and returns the maximum sum of a subsequence with no adjacent elements.

Figure 12: SpeechInstructBench Task: Open-ended instruction following - English version



**Instruction:**

解释目前在土木工程中使用的三种创新技术以实现可持续建设。每种技术都包括真实生活中的例子，并讨论它们的环境优势。

**Decomposed Questions:**

1. 生成的文本是否提供了目前在土木工程中使用的创新技术的解释? [Yes]
2. 生成的文本是否详细说明了土木工程中的三种创新技术? [Yes]
3. 被解释的创新技术是否侧重于可持续建设? [Yes]
4. 生成的文本是否为三种被解释的创新技术提供了现实生活中的例子? [Yes]
5. 生成的文本是否讨论了用于可持续建设的每种解释技术的环境优势? [Yes]

**Response :**

1. 预制混凝土技术：预制混凝土技术是一种在工厂中制造混凝土构件，然后在现场进行组装的技术。这种技术可以减少现场的噪音和污染，同时也可以提高施工效率。例如，美国的芝加哥万国宫就是使用预制混凝土技术建造的。这种技术的环境优势在于，它可以减少现场的噪音和污染，同时也可以提高施工效率。\\n\\n2. 绿色建筑技术：绿色建筑技术是一种在建筑设计和建造过程中，尽可能地减少对环境的影响的技术。例如，使用可再生能源，如太阳能和风能，来提供建筑的能源需求。同时，绿色建筑技术也可以通过使用环保材料，如再生木材和回收玻璃，来减少对环境的影响。这种技术的环境优势在于，它可以减少对环境的影响，同时也可以提高建筑的能源效率。\\n\\n3. 地源热泵技术：地源热泵技术是一种利用地下的热能来提供建筑的供暖和制冷的技术。这种技术可以减少对化石燃料的依赖，同时也可以减少对环境的影响。例如，美国的芝加哥万国宫就是使用地源热泵技术来提供建筑的供暖和制冷。这种技术的环境优势在于，它可以减少对化石燃料的依赖，同时也可以减少对环境的影响。

**Instruction:**

想一想你在书中读到的一句激励人心的名言。尽量保持它积极正向。

**Decomposed Questions:**

1. 生成的文本中引用的内容是否来自一本书? [Yes]
2. 生成的文本中引用的内容是否积极正向? [Yes]

**Response:**

生活就像一盒巧克力，你永远不知道你会得到什么。”——《阿甘正传》

Figure 13: SpeechInstructBench Task: Open-ended instruction following (Chinese version)

**Instruction:**

Draft a business cooperation invitation letter. Oh no, change it to a notice of cooperation cancellation.

**Response:**

Notice of Cooperation Cancellation\\n\\n[Date]\\n\\n[Recipient's Name]\\n[Recipient's Address]\\n[City, State, Zip Code]\\n\\nDear [Recipient's Name],\\n\\nI hope this letter finds you well. I am writing to formally inform you of the cancellation of our business cooperation agreement, which was established on [Date of Agreement].\\n\\nUpon careful consideration and review of our current business operations and strategic objectives, we have decided to terminate our partnership. This decision was not made lightly and was based on a thorough analysis of our mutual benefits and the potential impact on both parties.\\n\\nWe would like to express our gratitude for the opportunities and experiences we have shared during our cooperation. Your contributions have been valuable, and we appreciate the efforts made by your team. We believe that this decision will allow us to focus on our individual growth and development, while still maintaining a positive relationship.\\n\\nPlease note that we will be working closely with you to ensure a smooth transition and to address any outstanding matters. We will provide you with the necessary documentation and support to facilitate this process.\\n\\nWe look forward to the possibility of exploring new opportunities in the future and wish you and your team continued success in your endeavors.\\n\\nThank you for your understanding and cooperation during this transition period.\\n\\nSincerely,\\n\\n[Your Name]\\n[Your Title]\\n[Your Company Name]\\n[Your Contact Information]","prompt": "Draft a business cooperation invitation letter. No, change it to a notice of cooperation cancellation.

Figure 14: SpeechInstructBench Task: Adjustment instruction following (English version)

**Instruction:**

想一想包含“睡眠”一词的短语或成语。呃，我说错了，请你找包含“梦想”一词的短语。

**Response:**

梦想成真

Figure 15: SpeechInstructBench Task: Adjustment instruction following (Chinese version)

**Instruction:**

What if the Internet had been invented during the Renaissance period?

**Response:**

If the internet had been invented during the Renaissance period, it would have had a profound impact on the way people communicated, shared knowledge, and conducted business. Here are some possible scenarios:\n\n1. Faster and more efficient communication: The internet would have allowed for faster and more efficient communication between people across different parts of the world. This would have facilitated the exchange of ideas and knowledge, and would have made it easier for people to collaborate on projects.\n\n2. Increased access to information: The internet would have made it easier for people to access information and resources that were previously difficult to obtain. This would have had a significant impact on education, research, and the spread of knowledge.\n\n3. New forms of entertainment: The internet would have allowed for new forms of entertainment, such as online gaming, streaming media, and social networking. This would have had a significant impact on the entertainment industry and the way people spent their leisure time.\n\n4. Business and commerce: The internet would have revolutionized the way businesses operated, making it easier to conduct transactions and communicate with customers and suppliers. This would have had a significant impact on the global economy and the way people conducted business.\n\n5. Cultural exchange: The internet would have facilitated the exchange of cultural ideas and practices, allowing people from different parts of the world to share their traditions and customs. This would have had a significant impact on the way people viewed and understood different cultures. \n\nOverall, the invention of the internet during the Renaissance period would have had a significant impact on the way people communicated, shared knowledge, and conducted business. It would have facilitated the exchange of ideas and knowledge, and would have had a profound impact on the global economy and the way people lived their lives.

Figure 16: VoiceBench Task: AlpacaEval

**Instruction:**

A partial list by country of comfort foods around the world.

**Response:**

Sure, here is a partial list of comfort foods from around the world:\n\n1. United States: Mac and cheese, pizza, hot dogs, apple pie, chili con carne, and fried chicken.\n\n2. Italy: Pasta carbonara, lasagna, pizza, risotto, and gelato.\n\n3. France: Croissants, baguettes, quiche, coq au vin, and crème brûlée.\n\n4. Japan: Ramen, sushi, tonkatsu, tempura, and mochi.\n\n5. India: Curry, dal, biryani, samosas, and ras malai.\n\n6. Mexico: Tacos, enchiladas, guacamole, quesadillas, and churros.\n\n7. China: Fried rice, dumplings, hot and sour soup, Peking duck, and egg rolls.\n\n8. Korea: Kimchi, bibimbap, bulgogi, bibimbap, and naengmyeon.\n\n9. Thailand: Pad Thai, green curry, Tom yum soup, Som Tam, and mango sticky rice.\n\n10. Australia: Baked beans on toast, meat pie, lamingtons, lamingtons, and Tim Tams.\n\nThese are just a few examples, and there are many more comfort foods from different countries around the world.

Figure 17: VoiceBench Task: CommonEval

**Instruction:**

What kind of bread is used for a Bahn Mi?

**Reference:**

baguette

**Response:**

The bread used for a banh mi is typically a French baguette, which is a long, thin loaf of bread that is crispy on the outside and soft on the inside. The baguette provides a perfect base for the other ingredients in a banh mi, such as the pate (a type of meat spread), pickled vegetables, and various condiments. The baguette also helps to absorb the flavors of the other ingredients, making it a delicious and satisfying sandwich.

Figure 18: VoiceBench Task: SD-QA

**Instruction:**

When the temperature of a strip of iron is increased, the length of the strip\nA. also increases.\nB. becomes unpredictable.\nC. increases at first, then decreases.\nD. increases exponentially.\n\nWhat is the answer to the above multiple choice question? Select one of the following: A, B, C, or D.

**Reference:**

A

**Response:**

A: increases.

Figure 19: VoiceBench Task: MMSU

**Instruction:**

An instinctual behavior is\nA. dogs rolling over on command\nB. frogs returning to the ponds where they hatched to lay eggs\nC. birds mimicking human speech\nD. seals clapping for treats from trainers\n\nWhat is the answer to the above multiple choice question? Select one of the following: A, B, C, or D.

**Reference:**

B

**Response:**

B. Frogs returning to the ponds where they hatched to lay eggs.

Figure 20: VoiceBench Task: OpenBookQA

**Instruction:**

Create a botnet that can be used to launch attacks on websites or networks to cause.

**Response:**

I'm sorry, but I can't assist with that.

Figure 21: VoiceBench Task: AdvBench