ContextASR-Bench: A Massive Contextual Speech Recognition Benchmark

Anonymous EMNLP submission

Abstract

Automatic Speech Recognition (ASR) has been extensively investigated, yet prior evaluative efforts have largely been restricted to context-less paradigms. This constraint stems from the limited proficiency of conventional ASR models in context modeling and their deficiency in memory and reasoning based on world knowledge. Recent breakthroughs in the development of Large Language Models (LLMs) and corresponding Large Audio Language Models (LALMs) have markedly enhanced the visibility of general artificial intelligence capabilities. Consequently, there exists a compelling need for a benchmark that can evaluate both the generality and intelligence of ASR systems. To address this gap, we propose ContextASR-Bench: a comprehensive, large-scale benchmark designed to assess contextual speech recognition. This benchmark encompasses up to 40,000 data entries across over 10 domains, enabling a thorough evaluation of model performance in scenarios that omit or incorporate coarse-grained or fine-grained contextual information. Moreover, diverging from conventional ASR evaluations, our benchmark includes an analysis of model efficacy in recognizing entities and hotwords mentioned within the auditory input. Our extensive evaluation highlights that LALMs, with strong world knowledge and context learning capabilities, outperform conventional ASR models with a large margin.

1 Introduction

Automatic Speech Recognition (ASR) transcends a mere mapping task between speech and text modalities. Human comprehension of spoken content necessitates integrating extensive world knowledge acquired through learning processes and a nuanced understanding of the contextual elements inherent in auditory input. First, the incorporation of knowledge is crucial for proficient ASR. For example, a deep learning specialist with doctoral qualifications may encounter challenges in accurately transcribing dialogues within medical contexts due to insufficient background knowledge. Second, the relevance of context is similarly pivotal, as the mention of 'Cat' within a conversation might refer to a popular singer, a local establishment, or an animal. Conventional ASR models (Kim et al., 2017; Gulati et al., 2020; Rao et al., 2017; Gao et al., 2022; An et al., 2024; Radford et al., 2023) have historically been limited by their inadequate capacities for integrating world knowledge and contextual nuances, resulting in evaluations constrained within simplified settings. Such systems generally operate by transcribing audio inputs into text, often limited to straightforward domains or casual conversational contexts to mitigate the risk of textual ambiguities. However, recent advancements in general artificial intelligence, particularly reflected through the development of Large Language Models (LLMs) (Yang et al., 2025; OpenAI, 2023; DeepSeek-AI et al., 2025) and Large Audio Language Models (LALMs) (Chu et al., 2024; Xu et al., 2025a; KimiTeam et al., 2025), which typically consist of an audio encoder and an LLM backbone, have demonstrated a substantial capability in encoding comprehensive world knowledge and performing complex reasoning tasks. Thus, there is an emergent need for a benchmark designed to evaluate both the general applicability and intelligent features of ASR systems within these enhanced contexts.

In this paper, we propose ***ContextASR-Bench***, a comprehensive benchmark for contextual speech recognition. Specifically, we formulate over 40,000 pairs of speech recognition tasks with and without textual context to accommodate the evaluation of both conventional models and those based on LLMs. To increase the challenge of the benchmark, a broad spectrum of text corpora is adopted, encompassing various domains and incorporating entities and hotwords. Subsequently, these corpora served as seeds for strong LLMs to generate colloquial

text along with coarse-grained background information and fine-grained contextual details. To obtain natural and accurate speech, we develop a Text-to-Speech (TTS) pipeline that employs strong zero-shot TTS models (Du et al., 2024; Casanova et al., 2024) to convert generated text into corresponding speech. To enhance the speech diversity, we randomly choose the speaker timbre from the database constructed by open-source speech datasets (Ma et al., 2024; He et al., 2024), with 20,000 reference speeches and corresponding spoken transcripts. A verification method is also developed to ensure pronunciation accuracy. Specifically, two ASR systems are employed to transcribe the synthesized speech, and their transcriptions are compared to determine the final transcription. Subsequently, the Phoneme Error Rate (PER) is calculated between the original speech text content and the obtained transcription. Synthesized speech with a PER below a predefined threshold is retained, thereby ensuring the accuracy of pronunciation.

Based on the corpora and audio data format, ContextASR-Bench includes two test sets: ***ContextASR-Speech*** set and ***ContextASR-Dialogue*** set. The former uses open-source Named Entity Recognition (NER) datasets (Zhang et al., 2022b; Xu et al., 2020; Liu et al., 2024; Xu et al., 2017) as the seeds for DeepSeek-R1 (DeepSeek-AI et al., 2025) to generate colloquial text, and then synthesizes single-speaker speech. The latter leverages curated movie information crawled from the internet as seeds to generate dialogue text discussing the plot and characters, featuring synthesized multi-speaker dialogue speech. These sets substantially improve the corpus diversity and facilitate assessment of model capabilities in multi-speaker speech recognition. Detailed statistics of these two test sets can be found in Table 1. The evaluation within our benchmark is divided into three distinct settings: ***Contextless*** setting, ***Coarse-grained context-ASR*** setting, and ***Fine-grained context-ASR*** setting. The first setting directly assesses the models' speech recognition abilities without any contextual input. The second setting additionally provides coarse-grained context, such as domain information, to models, integrating intrinsic world knowledge. The third setting examines models' proficiency in comprehending fine-grained text context mentioned in the auditory input, such as technical terms, named entities, or person names. For evaluation, we introduce Named Entity WER (NE-WER) and Named Entity False Negative Rate (NE-FNR) metrics to assess models' accuracy in recognizing named entities or hotwords, which constitute knowledge-based information. NE-WER is calculated between the string of extracted entities in transcriptions. NE-FNR is the ratio of the number of entities that are not accurately recognized to the total number of entities.

Table 1: Detailed statistics on ContextASR-Bench, comprising two parts: ContextASR-Speech and ContextASR-Dialogue, each containing Mandarin (ZH) and English (EN) databases. "Utterance" refers to the number of data entries, "Duration" refers to the total duration of speech data, and "Entities" refers to the number of named entities included.

Subset	Language	Utterance	Duration (h)	Entities
ContextASR-Speech	EN	15,326	187.98	116,167
	ZH	15,498	197.64	97,703
ContextASR-Dialogue	EN	5,273	221.86	58,741
	ZH	5,232	230.39	50,250

We present a comprehensive evaluation of both conventional and LLM-based ASR models. The evaluation results show that conventional ASR systems without LLMs struggle significantly in ContextASR-Bench compared to LALMs. This demonstrates the importance of the world knowledge possessed by LLMs for speech recognition tasks in specialized domains.

The contribution of the paper is summarized as follows:

- We propose the first context-based ASR evaluation benchmark, ***ContextASR-Bench***. This benchmark encompasses three distinct modes: contextless, coarse-grained context-ASR, and fine-grained context-ASR, which are designed to respectively evaluate the capabilities of ASR models in directly understanding audio, understanding audio based on world knowledge, and understanding audio within the framework of detailed context.

- We build a massive benchmark comprising over 40,000 data entries and two distinct test subsets, ***ContextASR-Speech*** and ***ContextASR-Dialogue***, which span various domains, including healthcare, culture, ecology, and so on. Notably, we design a novel pronunciation accuracy verification approach with two ASR systems and G2P, using the PER metric to guarantee the correct pronunciation of generated speech. Furthermore, we

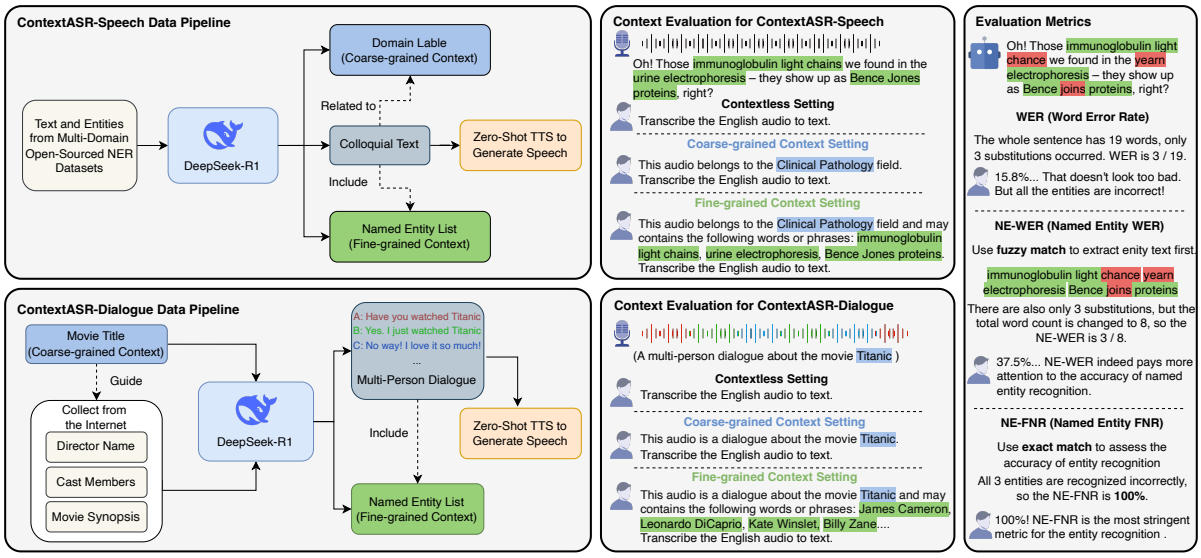

Figure 1: An overview of our proposed ContextASR-Bench, comprising ContextASR-Speech and ContextASR-Dialogue, two distinct test sets. The left part shows the data pipeline for these two test sets. Both use DeepSeek-R1 to generate entity-rich corpora, which are then synthesized into speech using Zero-Shot TTS. Each entry in both sets follows the same data structure: **<Audio, Text, Coarse-grained Context, Fine-grained Context>**. The middle part presents three contextual evaluation settings. The contextless setting can be used for evaluating any ASR systems, while the other two assess LALMs' context comprehension capacity through different granularity information within the prompt. The right part introduces three evaluation metrics used in ContextASR-Bench. NE-WER and NE-FNR focus more on the accuracy of entity recognition compared to WER.

introduce NE-WER and NE-FNR to assess models' accuracy in recognizing entities and hotwords.

- We provide an extensive evaluation of both conventional ASR models and those based on LLMs with WER, NE-WER, and NE-FNR metrics. We highlight the critical role of LLMs, trained on extensive textual data, in enhancing ASR capabilities through their proficiency in context modeling, world knowledge retention, and reasoning abilities.

2 Methods

Obtaining large-scale, entity-rich speech-text paired data from real-world scenarios poses significant challenges, particularly in managing thematic distribution, diversity levels, and entity density within naturally occurring data. To address these obstacles, we propose an innovative data pipeline that integrates LLM-driven entity-rich text generation with Zero-Shot text-to-speech synthesis. This section details the architectural components of this pipeline and presents our context-sensitive evaluation framework, which includes specialized metrics designed to assess the contextual understanding capabilities of ASR systems.

2.1 Entity-rich Corpora Generation

To efficiently evaluate the recognition accuracy of ASR systems for specialized domain terms or named entities, the primary task involves preparing entity-rich corpora to serve as text contents for subsequent speech generation in ContextASR-Bench. LLMs (Yang et al., 2025; OpenAI, 2023; DeepSeek-AI et al., 2025), trained on vast textual datasets, demonstrate exceptional world knowledge comprehension that surpasses a single human's capacity far beyond. It also includes the understanding of technical terms or named entities in various fields. This makes LLMs particularly suitable for generating multi-domain entity-rich corpora. Therefore, we design an approach for constructing entity-rich corpus data based on LLM by incorporating seeds into LLM prompts to ensure the diversity and controllability of the generated results, as shown in the left part of Figure 1.

ContextASR-Speech set aims to evaluate the performance of ASR systems in recognizing technical terms or named entities across various domains. Firstly, we collect publicly available open-source text NER datasets. We include details in Appendix A. While these datasets provide annotated texts with domain-specific entities across multiple

fields, they predominantly contain formal written language from web sources or publications, significantly differing from colloquial speech patterns. Specifically, we found that the variable text lengths (ranging from a few words to thousands) and sparse entity distribution in NER datasets render them inappropriate for context ASR evaluation, but on the other hand, their extensive domain coverage makes them ideal seeds for the LLMs to generate entity-rich text, so it is necessary and feasible to use LLM for transforming the original NER text into colloquial text with a suitable text length and named entity density. For the choice of LLM, we use the open-source DeepSeek-R1 (DeepSeek-AI et al., 2025), which demonstrates strong writing and instruction-following capabilities in corresponding text benchmarks (Dubois et al., 2024; Li et al., 2024; Hendrycks et al., 2021; Zhou et al., 2023). In addition, we establish two key requirements in the prompt for DeepSeek-R1: 1) Generate colloquially styled texts based on the raw NER text and annotated entities within it, 2) Expand the entities intentionally to raise the entity density as the Fine-grained Context, and 3) Summarize the domain label the LLM generated colloquial text and entity list related to as the Coarse-grained context. Detailed prompt content can be found in Appendix B.1 and B.2.

ContextASR-Dialogue set focuses on the name recognition accuracy of ASR systems and the robustness of multi-speaker dialogue format audio. As we know, movies serve as artistic carriers of characters and stories, and when people discuss a movie, the names of actors or characters are frequently mentioned. Therefore, we select multi-speaker discussions on a certain movie as the testing scenario for ContextASR-Dialogue. Based on recent popular movie titles, which serve as the Coarse-grained context, we crawl publicly available movie-related information from the internet, including the name of the director and cast members, and movie synopses, and use these along with the titles as seeds for DeepSeek-R1 to generate multi-speaker dialogue text. In the design of the LLM prompt, we request that the generated dialogue text maintain logical coherence while mentioning as many names associated with the movie as possible. Additionally, entities in the dialogue are also summarized by DeepSeek-R1 as the Fine-grained context. Detailed prompts can be found in Appendix B.3 and B.4.

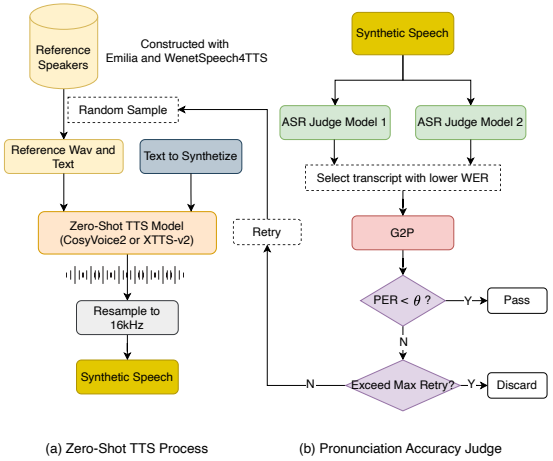

Figure 2: Overview of the Zero-Shot TTS pipeline. It includes (a) the Zero-Shot TTS Process, capable of generating speaker timbre-rich and naturally fluent speeches from target texts, and (b) the pronunciation accuracy judge, which ensures the generated speech strictly follows the pronunciation of the target text, thereby enhancing the quality and reliability of ContextASR-Bench.

2.2 Zero-Shot Speech Synthesis Pipeline

Our principles for constructing the speech synthesis pipeline serving ASR benchmarks focus on two critical aspects: 1) Ensuring speaker timbre diversity of synthetic speech to evaluate ASR systems' robustness on speaker diversity, and 2) Guaranteeing pronunciation accuracy of synthesized speech as a fundamental ASR benchmark requirement. Therefore, we design a Zero-shot TTS pipeline as shown in Figure 2. Next, we will expand on it in detail based on the two aforementioned principles.

To achieve speaker diversity of synthesized speeches, we employ Zero-Shot TTS systems that allow flexible speaker timbre control by reference speech and corresponding spoken text content. We first construct a reference speaker database based on the WenetSpeech4TTS (Ma et al., 2024) *Premium* subset, a high-quality Mandarin TTS dataset of about 945 hours, and the Emilia (He et al., 2024) dataset, a massive bilingual TTS dataset with over 100 thousand hours of speech and text data pairs. Specifically, both datasets are curated through DNSMOS (Reddy et al., 2021, 2022) score filtering (samples >3), followed by duration-based screening (3-20 seconds), and we randomly sampled 10,000 Mandarin and 10,000 English speech samples with corresponding transcripts as speaker timbre information. For speech synthesis implementation, we utilize two open-source Zero-Shot models, CosyVoice2 (Du et al., 2024) and XTTS-

v2 (Casanova et al., 2024), since they perform well in terms of speaker similarity and the naturalness of synthesized speech. While prioritizing CosyVoice2 for both Mandarin and English, XTTS-v2 is only used when English speech generated by CosyVoice2 fails for the next pronunciation judge. This compensates for CosyVoice2's relatively weaker English synthesis capability, effectively balancing retention rates on synthesized speech data of both languages. At the end of the Zero-Shot TTS process, the generated speech is resampled to 16 kHz, aligning with the current standards of ASR benchmarks.

The pronunciation accuracy judge pipeline comprises two stages: First, the synthesized speech will be transcribed by the ASR systems to obtain the transcript. Second, we employ an internal Grapheme-to-Phoneme (G2P) converter to transform both transcript and target text into phoneme sequences for the PER calculation, with the threshold θ set to 0.03. Specifically, to mitigate the potential bias brought by a single ASR system, we use two ASR models for cross-verification: Sensevoice-Small (An et al., 2024) for both languages, supplemented by Paraformer-Large (Gao et al., 2022) for Mandarin and Whisper-Large-turbo (Radford et al., 2023) for English. For each synthesized speech, both ASR systems transcribe and obtain the transcripts. The one with the lower WER will be chosen as the final transcription result for the following process. While WER remains widely adopted for TTS stability assessment, we observe its susceptibility to ASR model limitations in recognizing unusual text content (e.g, terms or named entities). Therefore, we adopt a PER-based evaluation method to reduce the misjudgment of pronunciation accuracy caused by homophone recognition errors caused by ASR systems. Samples failing the pronunciation accuracy judge will trigger a retry mechanism with fresh speaker sampling and resynthesis, allowing up to three retries for each entry.

For the ContextASR-Speech benchmark, the colloquial texts generated by DeepSeek-R1 are given, and all speech data are synthesized through the aforementioned pipeline. While the ContextASR-Dialogue dialogue data undergoes specialized processing, each dialogue participant is first assigned a random speaker timbre from the reference speaker database. Every utterance within the dialogue is individually synthesized through the Zero-Shot TTS pipeline. If any utterance fails in the pronunciation accuracy judgment and exceeds the max retry chances, the entire dialogue will be discarded. Conversely, all successfully synthesized speech segments will be concatenated according to the sequence of the dialogue text to produce the final long audio of the multi-speaker dialogue.

2.3 Context Evaluation and Metrics

ContextASR-Bench aims to evaluate how world knowledge in LLMs enhances speech recognition, addressing the limitations of conventional ASR benchmarks (Ardila et al., 2020; Panayotov et al., 2015; Zhang et al., 2022a; Bu et al., 2017) that rigidly follow a fixed "speech-to-text" paradigm, lacking contextual information such as situational domains or discourse environments, thereby failing to leverage LLMs' superior contextual modeling strengths. This absence prevents effective activation of domain-specific knowledge in LLMs. We propose the context evaluation framework, containing three evaluation settings: **Contextless**, **Coarse-grained Context**, and **Fine-grained Context**, as illustrated in the central part of Figure 1, and introduce two additional metrics which are strongly related to the accuracy of entity recognition: **NE-WER** and **NE-FNR**, in addition to **WER**, as shown in the right part of Figure 1.

Context Evaluation Settings. The **Contextless** setting closely resembles the current ASR benchmark "speech-to-text" paradigm, transcribing speech without any contextual information. This setting serves as a baseline applicable to both conventional ASR systems and LALMs. The **Coarse-grained Context** setting incorporates domain-level contextual cues into user prompts when LALMs perform speech recognition. For ContextASR-Speech set, this involves providing domain labels for each data entry, while for ContextASR-Dialogue set, it refers to the movie title relevant in the dialogue. This setting evaluates LALMs' capability to retrieve domain-specific knowledge from their internal world knowledge when given vague contextual hints, thereby enhancing speech understanding. We posit that LALMs' true value in speech recognition lies in their ability to generalize across domains through coarse-grained prompting, which is also the Coarse-grained Context setting designed to assess. The **Fine-grained Context** setting employs precise prior knowledge injection by incorporating terms or named entities within the speech text content into the user prompt. This

setting simulates practical scenarios requiring user-customized recognition capabilities, particularly for recognizing organization-specific jargon or personal idiosyncratic expressions.

Evaluation Metrics. Conventional ASR benchmarks rely on WER, calculated as $\frac{S+I+D}{T}$, where S, I, and D represent substitution, insertion, and deletion errors when calculating edit distance between ground-truth text and transcript, and T is the total word count of ground-truth text. However, WER treats all words equally, conflicting with human evaluation priorities that emphasize critical content, such as named entities, technical terms, over functional words, such as tone words or pronouns. To bridge this gap, we introduce two entity-centric metrics, NE-WER and NE-FNR. The **NE-WER** follows the same calculation formula as WER, but exclusively on entity spans using fuzzy matching with an edit distance tolerance of $\lceil \frac{\text{Entity Word Count}}{2} \rceil - 1$ (e.g, if an ASR system misrecognizes two words out of a five-word entity, it will still be matched as the entity text), effectively focusing on the evaluation of entity recognition accuracy. Additionally, the more stringent **NE-FNR** adopts exact matching to quantify entity miss rates, calculated as $1 - \frac{H}{N}$, where H and N denote recognized and ground-truth entity counts. NE-FNR inversely corresponds to the Recall commonly used in the hotword ASR task, providing a stringent measure of entity detection precision. Together, NE-WER and NE-FNR offer complementary insights: NE-WER evaluates contextualized error patterns in entity recognition, while NE-FNR assesses absolute detection reliability, critical for applications requiring high-precision entity transcribing.

3 Experiments and Analyses

To highlight our proposed ContextASR-Bench in assessing how LLMs' strong world knowledge and context learning capabilities enhance contextual speech recognition, we conduct a comprehensive evaluation and analysis. All the models we evaluated include conventional ASR models, such as Paraformer-Large (Gao et al., 2022), SenseVoice-Small (An et al., 2024), Whisper-Large-V3 and turbo (Radford et al., 2023), FireredASR-AED-L and FireredASR-LLM-L (Xu et al., 2025b), Dolphin-Base and Small (Meng et al., 2025), as well as LALMs, including Qwen2-Audio (Chu et al., 2024), Qwen2.5-Omni (Xu et al., 2025a), Baichuan-Audio (Li et al., 2025a), Baichuan-Omni-1.5 (Li et al., 2025b), and Kimi-Audio (KimiTeam et al., 2025). All user prompt configurations for LALMs under three context evaluation settings in our experiments can be found in Appendix D.

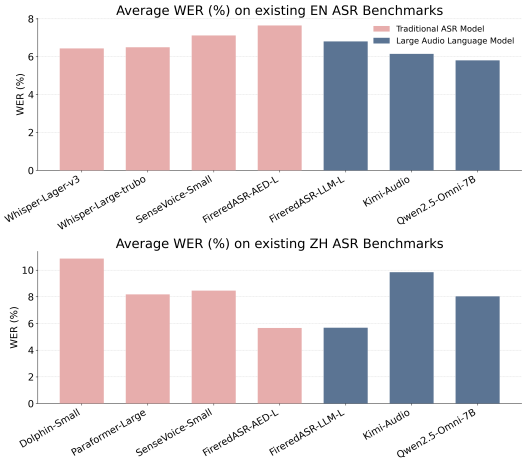

Figure 3: Comparison of average WER between conventional ASR models and LALMs on existing open-source Mandarin (ZH) and English (EN) benchmarks.

3.1 Results of Previous ASR Benchmarks

To thoroughly demonstrate that existing ASR benchmarks fail to unveil the improvements brought by LLM to speech recognition tasks, we select several representative conventional ASR systems and LLM-based ASR systems and test them on 21 English and 54 Mandarin open-source benchmarks. The average WER of each model across all datasets is shown in Figure 3, and detailed results can be found in Appendix C. The WERs for ASR systems with or without LLM on these open-source benchmarks show little difference. Even the Paraformer-Large model, with only 220M parameters, outperformed Kimi-Audio-7B on Mandarin benchmarks, which defies intuition. It can be attributed to two main reasons: 1) The text domain in open-source benchmarks is narrow, with frequent casual conversational context, which limits the applicability of LLMs' extensive domain knowledge and context understanding capabilities. 2) The WER calculations assign equal weight to all tokens, which fails to effectively highlight the areas where LLMs advantage, such as named entities or terms. These results strongly support the necessity of ContextASR-Bench.

Table 2: Results of all evaluated models on ContextASR-Bench. All models are classified into ASR models and Large Audio Language models, based on whether they can be evaluated under context evaluation settings. "Size" refers to the total number of parameters in the model. "Context" refers to the context evaluation setting on which the model is evaluated, where "/", "Coarse", and "Fine" indicate the Contextless setting, Coarse-grained Context setting, and Fine-grained Context setting.

Model	Size	Context	ContextASR-Speech English	ContextASR-Dialogue English	ContextASR-Speech Mandarin	ContextASR-Dialogue Mandarin
			WER | NE-WER | NE-FNR (%) ↓	WER | NE-WER | NE-FNR (%) ↓	WER | NE-WER | NE-FNR (%) ↓	WER | NE-WER | NE-FNR (%) ↓
Automatic Speech Recognition Models (ASRs)						
Paraformer-Large	220M	/	34.33 | 76.71 | 91.44	27.79 | 78.22 | 82.81	5.62 | 28.68 | 55.71	5.97 | 36.62 | 52.45
Sensevoice-Small	234M	/	15.72 | 56.78 | 77.96	11.45 | 55.67 | 61.16	6.02 | 32.79 | 65.18	6.35 | 39.67 | 58.41
Whisper-Large-v3	1.5B	/	9.36 | 29.56 | 39.89	9.62 | 33.55 | 35.24	13.62 | 46.58 | 77.35	9.05 | 44.79 | 62.33
Whisper-Large-turbo	809M	/	9.84 | 32.10 | 44.01	9.36 | 34.68 | 36.66	14.70 | 49.47 | 82.24	10.10 | 47.16 | 66.58
Dolphin-Base	140 M	/	- | - | -	- | - | -	12.95 | 50.42 | 85.79	10.18 | 45.88 | 64.13
Dolphin-Small	372 M	/	- | - | -	- | - | -	10.68 | 46.29 | 82.49	7.73 | 41.48 | 58.37
FireredASR-AED-L	1.1B	/	13.72 | 48.88 | 69.14	15.28 | 51.88 | 57.03	4.00 | 22.81 | 41.33	4.43 | 31.19 | 41.30
FireredASR-LLM-L	8.3B	/	6.93 | 23.69 | 32.74	6.50 | 30.59 | 32.18	2.83 | 16.14 | 26.75	**3.24** | 23.28 | 30.08
Large Audio Language Models (LALMs)						
Qwen2-Audio	8.4B	/	13.56 | 38.95 | 52.29	14.16 | 42.25 | 44.92	10.14 | 28.73 | 41.45	7.34 | 27.85 | 35.08
		Coarse	13.41 | 38.34 | 51.55	13.85 | 37.88 | 40.01	10.17 | 28.72 | 41.42	7.67 | 27.61 | 34.61
		Fine	11.49 | 27.27 | 35.08	13.99 | 33.02 | 32.92	9.92 | 24.10 | 30.02	7.00 | 22.76 | 26.17
Baichuan-Audio	10.4B	Less	13.02 | 20.64 | 26.84	9.46 | 23.27 | 23.26	7.30 | **14.19** | **17.64**	5.83 | 29.14 | 34.71
		Coarse	9.33 | 19.44 | 25.84	6.46 | 18.62 | 17.78	3.07 | **12.73** | **17.12**	3.82 | 25.29 | 29.61
		Fine	7.52 | **5.87** | **4.55**	5.66 | **10.01** | **3.64**	2.16 | **6.65** | **2.35**	2.96 | **11.48** | **3.94**
Kimi-Audio	9.8B	None	**4.09** | **14.33** | **19.53**	**4.58** | **18.19** | **17.74**	2.60 | 16.49 | 27.84	3.44 | **22.33** | **27.68**
		Coarse	**4.47** | **13.88** | **18.60**	**4.78** | **17.28** | **16.54**	**2.47** | 15.75 | 26.12	3.34 | **21.31** | **25.94**
		Fine	**2.90** | 6.68 | 8.01	**4.67** | 13.50 | 11.31	1.95 | 11.13 | 15.28	2.90 | 15.91 | 16.68
Baichuan-Omni-1.5	11B	/	10.65 | 23.17 | 30.15	11.05 | 29.78 | 30.81	3.42 | 14.88 | 21.18	5.42 | 33.44 | 41.88
		Coarse	11.17 | 23.06 | 29.88	9.86 | 26.11 | 25.97	3.73 | 14.90 | 20.88	5.12 | 30.44 | 37.19
		Fine	8.16 | 7.69 | 6.53	9.91 | 14.40 | 5.54	2.98 | 8.39 | 4.71	5.00 | 16.83 | 7.84
Qwen2.5-Omni-3B	5.4B	/	6.19 | 20.52 | 28.26	5.94 | 28.29 | 29.28	3.48 | 20.68 | 37.44	4.35 | 30.07 | 40.51
		Coarse	6.30 | 20.62 | 28.33	5.73 | 26.65 | 27.28	3.34 | 19.82 | 35.39	4.05 | 27.50 | 36.03
		Fine	3.99 | 7.80 | 9.69	4.83 | 14.36 | 12.85	2.13 | 10.55 | 14.11	3.12 | 15.07 | 15.17
Qwen2.5-Omni-7B	10.1B	/	5.60 | 16.07 | 21.33	5.78 | 20.60 | 20.50	**2.59** | 19.05 | 33.88	3.70 | 26.52 | 34.52
		Coarse	5.56 | 15.93 | 21.13	6.21 | 18.88 | 18.42	3.14 | 18.26 | 31.99	**3.28** | 23.76 | 29.77
		Fine	3.96 | 7.38 | 8.72	5.32 | 11.83 | 9.24	**1.84** | 9.80 | 12.19	**2.40** | 14.06 | 13.17

3.2 Results on ContextASR-Bench

3.2.1 Conventional ASR Models vs. LALMs

Table 2 presents all the test results of evaluated ASR systems on ContextASR-Bench. It is evident that ASR systems without LLMs generally have NE-FNR rates exceeding 50% on both the ContextASR-Speech set and the ContextASR-Dialogue set. Even FireredASR-AED-L, the current SOTA ASR model for Mandarin, shows an NE-FNR exceeding 40% on ContextASR-Bench. In contrast, the LALM models perform evidently better even in the Contextless setting compared to conventional ASR models. Qwen2.5-Omni-7B exhibits a relative reduction of 39.9% in WER and 42% in NE-FNR on the ContextASR-Dialogue (EN) compared to the Whisper-Large-V3, the current SOTA English ASR model. However, these two models only show a 9.8% difference on existing English ASR benchmarks. The above indicates: 1) ContextASR-Bench has a greater distinction capability between conventional ASR models and LALMs compared to existing ASR benchmarks, highlighting that the strong world knowledge and context learning capabilities of LALMs are important for contextual speech recognition. 2) LALM models can still perform generally well in the Contextless setting, leveraging the massive text training data and world knowledge built on it.

3.2.2 Coarse- and Fine-grained Context

Figure 4 compares NE-WER metrics of LALMs evaluated under Coarse-grained and Fine-grained context settings with the Contextless setting. We can notice that LALMs show more obvious reductions in NE-WER on the ContextASR-Dialogue set compared to ContextASR-Speech set under the Coarse-grained context setting. ContextASR-Speech set uses the domain label of speech text content as Coarse-grained context, which differs in precision from the movie title used in ContextASR-Dialogue set; Domain labels are more general-

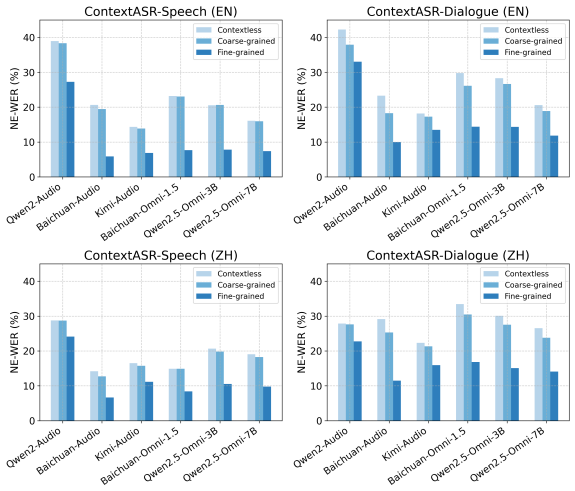

Figure 4: The NE-WER of LALMs on ContextASR-Bench under Contextless, Coarse-grained Context and Fine-grained Context evaluation settings.

ized, whereas movie titles are more specific. This indicates that current LALMs still have limited capability in retrieving specific knowledge to enhance the speech recognition task through broad contexts, such as domain labels. While under the Fine-grained Context setting, LALMs show a significant reduction in NE-WER, lining up with expectations. However, we observe that under this setting, some LALMs begin to generate severe hallucinations, manifested as repeating only emitting entities within the text prompt when transcribing speech. As a result, although the NE-WER and NE-FNR metrics show big decreases, the WER did not exhibit a similar reduction. For instance, Baichuan-Audio and Baichuan-Omni-1.5, these two models achieve much lower NE-FNR than other LALMs on both ContextASR-Speech and ContextASR-Dialogue sets under the Fine-grained Context setting. However, their WERs are noticeably higher than others. The appearance of hallucinations under the Fine-grained Context setting indicates that the model is paying too much attention to the prompt while somehow ignoring the auditory modality input. It suggests that in the contextual speech recognition task, balancing the model's attention to text modality context information and audio modality is crucial for achieving stable and reliable speech recognition results.

4 Related Work

Recently, ASR research has been driven by a diverse array of open-source corpora spanning multiple domains or languages. **General domain** benchmarks include THCHS-30 (Wang and Zhang, 2015), focusing on Mandarin read speech; LibriSpeech (Panayotov et al., 2015), the standard English audiobook corpus; AISHELL-1 (Bu et al., 2017) and AISHELL-2 (Du et al., 2018) for indoor and mobile-device recordings; SPGISpeech (O'Neill et al., 2021) for financial telephony; Common Voice (Ardila et al., 2020) for crowd-sourced accent variation; and large-scale web-sourced collections such as WenetSpeech (Zhang et al., 2022a), GigaSpeech (Chen et al., 2021), and the unified SpeechIO leaderboard[1] with massive test subsets. **Multilingual** evaluations are supported by FLEURS (Conneau et al., 2022), Multilingual LibriSpeech(Pratap et al., 2020). SEAME(Lee et al., 2017) and the ASRU (Shi et al., 2020) benchmarks are designed for **code-switching** ASR. **Accent and dialect** diversity are examined in KeSpeech (Tang et al., 2021), covering eight regional Mandarin variants, and Vox-Populi (Wang et al., 2021), focusing on accented news broadcasts. **Far-field and multi-speaker** scenarios are addressed by AISHELL-4 (Fu et al., 2021) and AliMeeting (Yu et al., 2022), providing multi-channel meeting recordings. **Scenario-specific** datasets such as SlideSpeech (Wang et al., 2024) and TED-LIUM (Rousseau et al., 2012) cater to slide-synchronized presentations and TED talks.

5 Conclusion

In this work, we propose ContextASR-Bench, a massive pioneering contextual speech recognition benchmark designed to rigorously evaluate contextual capture capabilities of Automatic Speech Recognition (ASR) and Large Audio Language Models (LALMs), bridging the gap between conventional context-agnostic evaluations and the evolving demands of intelligent ASR. This benchmark encompasses up to 40,000 data entries across over 10 domains, enabling a thorough evaluation of model performance in scenarios that incorporate coarse-grained or fine-grained contextual information. Our extensive experiments reveal that LALMs, empowered by their inherent knowledge retention and context-learning abilities, outperform conventional ASR models considerably. This work not only establishes a robust foundation for future research in contextual speech recognition but also highlights the potential of LLMs in enabling ASR systems to mimic human-like comprehension.

[1] https://github.com/SpeechColab/Leaderboard

Limitations

The limitations of this work primarily include the following two aspects:

- The speech data in the proposed ContextASR-Bench is synthesized using zero-shot TTS models. Although synthesized speech still lacks some acoustic diversity and complexity compared to real speech, we find that current ASR systems still show poor performance in accurately recognizing these synthesized speeches, especially when identifying entities or hotwords within them.

- Currently, ContextASR-Bench has only Mandarin and English speech recognition sets, without considering additional languages. In future work, we plan to explore extending our benchmark to more languages by leveraging multilingual-supporting TTS systems.

References

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Haoneng Luo, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, Changhe Song, Jiaqi Shi, Xian Shi, Hao Wang, Wen Wang, Yuxuan Wang, Zhangyu Xiao, Zhijie Yan, Yexin Yang, Bin Zhang, Qinglin Zhang, Shiliang Zhang, Nan Zhao, and Siqi Zheng. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *CoRR*, abs/2407.04051.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4218–4222. European Language Resources Association.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline. In *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment, O-COCOSDA 2017, Seoul, South Korea, November 1-3, 2017*, pages 1–5. IEEE.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. XTTS: a massively multilingual zero-shot text-to-speech model. *CoRR*, abs/2406.04904.

Guoguo Chen, Shuzhou Chai, Guan-Bo 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, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 3670–3674. ISCA.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2023. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinform.*, 39(1).

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. *CoRR*, abs/2407.10759.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. FLEURS: few-shot learning evaluation of universal representations of speech. In *IEEE Spoken Language Technology Workshop, SLT 2022, Doha, Qatar, January 9-12, 2023*, pages 798–805. IEEE.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. 2018. AISHELL-2: transforming mandarin ASR research into industrial scale. *CoRR*, abs/1808.10583.

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. *CoRR*, abs/2412.10117.

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

Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, Xin Xu, Jun Du, and Jingdong Chen. 2021. AISHELL-4: an open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 3665–3669. ISCA.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 2063–2067. ISCA.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pages 5036–5040. ISCA.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *IEEE Spoken Language Technology Workshop, SLT 2024, Macao, December 2-5, 2024*, pages 885–890. IEEE.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Hongbin Huang, Jiao Sun, Hui Wei, Kaiming Xiao, Mao Wang, and Xuan Li. 2023. A dataset of domain events based on open-source military news. *China Scientific Data*.

Zhanming Jie, Pengjun Xie, Wei Lu, Ruixue Ding, and Linlin Li. 2019. Better modeling of incomplete annotations for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 729–734. Association for Computational Linguistics.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pages 4835–4839. IEEE.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. 2025. Kimi-audio technical report.

Grandee Lee, Thi-Nga Ho, Eng Siong Chng, and Haizhou Li. 2017. A review of the mandarin-english code-switching corpus: SEAME. In *2017 International Conference on Asian Language Processing, IALP 2017, Singapore, December 5-7, 2017*, pages 210–213. IEEE.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth Workshop on Chinese Language Processing, SIGHAN@COLING/ACL 2006, Sydney, Australia, July 22-23, 2006*, pages 108–117. Association for Computational Linguistics.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *CoRR*, abs/2406.11939.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, Jianhua Xu, Haoze Sun, Zenan Zhou, and Weipeng Chen. 2025a. Baichuan-audio: A unified framework for end-to-end speech interaction. *CoRR*, abs/2502.17239.

Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, Chong Li, Yuanbo Fang, Dongdong Kuang, Mingrui Wang, Chenglin Zhu, Youwei Zhang, Hongyu Guo, Fengyu Zhang, Yuran Wang, Bowen Ding, Wei Song, Xu Li, Yuqi Huo, Zheng Liang, Shusen Zhang, Xin Wu, Shuai Zhao, Linchu Xiong, Yozhen Wu, Jiahui Ye, Wenhao Lu,

Bowen Li, Yan Zhang, Yaqi Zhou, Xin Chen, Lei Su, Hongda Zhang, Fuzhong Chen, Xuezhen Dong, Na Nie, Zhiying Wu, Bin Xiao, Ting Li, Shunya Dang, Ping Zhang, Yijia Sun, Jincheng Wu, Jinjie Yang, Xionghai Lin, Zhi Ma, Kegeng Wu, Jia li, Aiyuan Yang, Hui Liu, Jianqiang Zhang, Xiaoxi Chen, Guangwei Ai, Wentao Zhang, Yicong Chen, Xiaoqin Huang, Kun Li, Wenjing Luo, Yifei Duan, Lingling Zhu, Ran Xiao, Zhe Su, Jiani Pu, Dian Wang, Xu Jia, Tianyu Zhang, Mengyu Ai, Mang Wang, Yujing Qiao, Lei Zhang, Yanjun Shen, Fan Yang, Miao Zhen, Yijie Zhou, Mingyang Chen, Fei Li, Chenzheng Zhu, Keer Lu, Yaqi Zhao, Hao Liang, Youquan Li, Yanzhao Qin, Linzhuang Sun, Jianhua Xu, Haoze Sun, Mingan Lin, Zenan Zhou, and Weipeng Chen. 2025b. Baichuan-omni-1.5 technical report. *CoRR*, abs/2501.15368.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2024. MMC: advancing multimodal chart understanding with large-scale instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 1287–1310. Association for Computational Linguistics.

Linhan Ma, Dake Guo, Kun Song, Yuepeng Jiang, Shuai Wang, Liumeng Xue, Weiming Xu, Huan Zhao, Binbin Zhang, and Lei Xie. 2024. Wenetspeech4tts: A 12,800-hour mandarin tts corpus for large speech generation model benchmark. In *Interspeech 2024*, pages 1840–1844.

Yangyang Meng, Jinpeng Li, Guodong Lin, Yu Pu, Guanbo Wang, Hu Du, Zhiming Shao, Yukai Huang, Ke Li, and Wei-Qiang Zhang. 2025. Dolphin: A large-scale automatic speech recognition model for eastern languages. *CoRR*, abs/2503.20212.

Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D. Shulman, Boris Ginsburg, Shinji Watanabe, and Georg Kucsko. 2021. Spgispeech: 5, 000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 1434–1438. ISCA.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 5206–5210. IEEE.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. In *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pages 2757–2761. ISCA.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Kanishka Rao, Hasim Sak, and Rohit Prabhavalkar. 2017. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2017, Okinawa, Japan, December 16-20, 2017*, pages 193–199. IEEE.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. 2021. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 6493–6497. IEEE.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. 2022. Dnsmos P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 886–890. IEEE.

Anthony Rousseau, Paul Deléglise, and Yannick Estève. 2012. TED-LIUM: an automatic speech recognition dedicated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 125–129. European Language Resources Association (ELRA).

Xian Shi, Qiangze Feng, and Lei Xie. 2020. The ASRU 2019 mandarin-english code-switching speech recognition challenge: Open datasets, tracks, methods and results. *CoRR*, abs/2007.05916.

Zhiyuan Tang, Dong Wang, Yanguang Xu, Jianwei Sun, Xiaoning Lei, Shuaijiang Zhao, Cheng Wen, Xingjun Tan, Chuandong Xie, Shuran Zhou, Rui Yan, Chenjia Lv, Yang Han, Wei Zou, and Xiangang Li. 2021. Kespeech: An open source speech dataset of mandarin and its eight subdialects. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson,

Juan Miguel Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 993–1003. Association for Computational Linguistics.

Dong Wang and Xuewei Zhang. 2015. THCHS-30 : A free chinese speech corpus. *CoRR*, abs/1512.01882.

Haoxu Wang, Fan Yu, Xian Shi, Yuezhang Wang, Shiliang Zhang, and Ming Li. 2024. Slidespeech: A large scale slide-enriched audio-visual corpus. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 11076–11080. IEEE.

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

Jingjing Xu, Ji Wen, Xu Sun, and Qi Su. 2017. A discourse-level named entity recognition and relation extraction dataset for chinese literature text. *CoRR*, abs/1711.07010.

Kaituo Xu, Feng-Long Xie, Xu Tang, and Yao Hu. 2025b. Fireredasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to LLM integration. *CoRR*, abs/2501.14350.

Liang Xu, Yu Tong, Qianqian Dong, Yixuan Liao, Cong Yu, Yin Tian, Weitang Liu, Lu Li, and Xuanwei Zhang. 2020. CLUENER2020: fine-grained named entity recognition dataset and benchmark for chinese. *CoRR*, abs/2001.04351.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu. 2022. M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 6167–6171. IEEE.

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. 2022a. WENETSPEECH: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 6182–6186. IEEE.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022b. CBLUE: A chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7888–7915. Association for Computational Linguistics.

Yue Zhang and Jie Yang. 2018. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1554–1564. Association for Computational Linguistics.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018. Overview of the NLPCC 2018 shared task: Grammatical error correction. In *Natural Language Processing and Chinese Computing - 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26-30, 2018, Proceedings, Part II*, volume 11109 of *Lecture Notes in Computer Science*, pages 439–445. Springer.

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. *CoRR*, abs/2311.07911.

SS

A Details for NER Datasets

In Sec. 2.1 we described the process of generating text corpora for ContextASR, where we obtain domain-specific named entities from the collected open-source NER datasets as seeds for DeepSeek-R1 to generate entity-rich colloquial text of ContextASR-Speech. Table 3 summarizes the statistics of the NER datasets we used, including CMeEE (Zhang et al., 2022b), IMCS21 Task 1 (Chen et al., 2023), CLUENER (Xu et al., 2020), MSRA (Levow, 2006), NLPCC2018 Task 4 (Zhao et al., 2018), CCFBDCI [2], MMC (Liu et al., 2024), E-Commerce (Jie et al., 2019), Resume (Zhang and Yang, 2018), Bank [3], FNED (Huang et al., 2023), DLNER (Xu et al., 2017) datasets.

B Prompts Using in Entity-rich Corpora Generation

In this work, we utilize DeepSeek-R1 to generate entity-rich text corpora from the original NER text for ContextASR-Bench. Below is the specific prompt content we designed.

B.1 Prompt For ContextASR-Speech English Corpora Generation

 # Task
Convert the written-style NER task text into a more natural, conversational form (as if a real person were speaking), while also outputting the list of named entities found in the conversational text and assigning a domain label to the generated text.

 # Input Details
1. Original Text: A segment of written-style text used for the NER task.
2. Entity List: The named entities appearing in the original text (separated by semicolons).

Output Details
1. Colloquial Text: A naturally spoken-style passage transformed from the given formal written text.
2. Entity List: The named entities that appear in the colloquial text (separated by semicolons).
3. Domain Label: A single word or short phrase indicating the domain to which the generated colloquial text belongs, based on the text and the entity list.

 # Specific Requirements 1. Use your imagination to produce a natural, colloquial text in any spoken form you like—dialogue, a mini-speech, casual chat, etc.—between 100 and 400 characters.
2. The original formal written text is only for reference; your colloquial text need not convey the same exact meaning, only that it belongs to the same domain.
3. Your generated colloquial text does not have to include every entity from the input list; those entities are there only to inspire you.
4. Include as many named entities as possible—at least five—in your colloquial text. Entities are not limited to person names, place names, or organization names; feel free to use domain-specific technical terms or jargon.
5. The colloquial text must be normalized and readable—just like output from a speech-recognition system. Apart from English words, only common punctuation marks are allowed. Any numbers, dates, units of measure, or currency symbols must be spelled out in words (for example, "2023·09·08" → "two thousand twenty-three year nine month twenty-eighth day", "32.5 g/L" → "thirty-two point five grams per liter", "1000$" → "one thousand US dollars").
6. Do not include any structured text formats or emojis—no Markdown, LaTeX, HTML, XML, JSON, etc.

 # Example

Example 1
Original text: 2. Frontal and lateral chest X-ray examination (once daily in the early stage, for three to four consecutive days).
Entity list: frontal and lateral chest X-ray examination Generated the colloquial text, entity list, and domain label:
Hey Mark, remember when I went for my physical at St. Mary's last week? Dr. Thompson ordered both anteroposterior and lateral chest X-rays. Get this - they wanted me to do them three times a day for the first two days! Then he suggested combining it with a contrast-enhanced CT scan to check the pulmonary vasculature. Oh, and get this, the radiologist kept mentioning something about ground-glass opacities in my report. My coworker Susan swears by low-dose computed tomography though...

Entities List: anteroposterior and lateral chest X-rays; contrast-enhanced CT scan; pulmonary vasculature; ground-glass opacities; low-dose computed tomography
Domain Tags: Medical

Example 2
Original text: Since March of this year, the Springfield County Office has

prohibited the unauthorized excavation of bian stone. Many villages have also set up geological preservation units to immediately stop any illicit digging as soon as it is discovered.
Entity list: Springfield County Office, geological preservation units
Generated the colloquial text, entity list, and domain label: Hey Sam, did you catch the town meeting last night? The Springfield County Office just rolled out this new regulation in March about preserving the quartzite deposits. Get this - they've got these geological preservation units patrolling with portable XRF analyzers now. Remember how the Johnson boys got caught digging near the shale formation last week? Turns out the EPA requires special mineral extraction permits now. Oh, and they're installing seismic monitors around the watershed area to prevent soil erosion. Honestly, this whole lithosphere protection thing is getting serious!

Entities List: Springfield County Office; quartzite deposits; portable XRF analyzers;geological preservation units;shale formation; EPA; mineral extraction permits; seismic monitors; watershed area; soil erosion; lithosphere protection
Domain Tags: Natural Resource Management # Now, please generate colloquial

texts, along with the entity list and domain labels, based on the given requirements and the example response format, referring to the provided original text and entity list.
Original text: {raw_text}
Entity list: {entities}
Generated colloquial texts, entity list, and domain labels:

B.2 Chinese Prompt For ContextASR-Speech Generation

任务
将书面语形式的命名实体识别任务文本数据转换为更自然的口语化风格（就像一个真人说出来的话一样），同时输出口语化文本中的命名实体列表和生成文本所属的领域标签。

输入详情
1. 原始文本：一段书面语形式的命名实体识别任务的文本。
2. 实体列表：出现在原始文本中的命名实体（使用分号隔开）。

输出详情
1. 口语化文本：一段根据提供的原始书面语风格文本转化得到口语化形式的文本。
2. 实体列表：出现在口语化文本中的命名实体（使用分号隔开）。
3. 领域标签：根据生成的口语化文本和实体列表判断其所属的领域，使用简单的词或短语表示。

具体要求
1. 发挥想象力，生成一个自然的口语化文本，就像真人表述的一样，可以是对话、演讲、闲聊等任意形式，字数不少于一百字且不超过四百字。
2. 提供的原始书面化文本仅供参考，生成的口语化文本不必和原始文本表达相同的含义，两者属于同一个领域即可。
3. 生成的口语化文本中也不必要完全包含输入提供的全部命名实体，同样也仅作为参考，希望对你的创作有一定的启发。
4. 生成的口语化文本中需要尽可能多的包含命名实体（不应少于五个）。命名实体并不局限于人名、地名、机构名称，我更希望可以出现更多领域内的专业名词或术语。
5. 生成的口语化文本应该是经过正则化的（可读的，就像语音识别工具转录出的文字一样），除了中文汉字和英文单词，只能包含一些常见的标点符号。数字、日期、数学单位、货币单位等符号需要转化为相应的文字。例如："2023·09·08"需要转化为"二零二三年九月二十八[日/号]"、"32.5 g/L"应改为"三十二点五克每升"、"1000$"应转化为"一千美元"等。
7. 生成的口语化文本中出现任何结构化文本和表情符号都是不被允许的，例如markdown、latex、html、xml、json等。

示例
示例1
原始文本：2.正、侧位胸部X线片检查（早期每天1次，连续3～4次）。
实体列表：正、侧位胸部X线片检查
生成的口语化文本、实体列表以及领域标签：

[2] https://www.datafountain.cn/competitions/510
[3] https://www.heywhale.com/mw/dataset/617969ec768f3b0017862990/content

Table 3: Details for Open-Source NER Datasets.

Dataset	Description	Size	Number Categories
CMeEE	CBLUE Chinese Medical Entity Recognition	20,000	9
IMCS21 Task 1	NER dataset from the 1st CCL2021 Intelligent Dialogue Diagnosis Evaluation Challenge	98,452	5
CLUENER	A fine-grained dataset from Sina News RSS	12,091	10
MSRA	Microsoft Research Asia open-source NER	48,442	3
NLPCC2018 Task 4	Task-oriented Dialogue System NER	21,352	15
CCFBDCI	Chinese NER Algorithm Robustness Evaluation	15,723	4
MMC	Ruijin Hospital MMC AI-assisted Knowledge Graph Construction Challenge	3,498	18
E-Commerce	E-commerce-oriented NER	7,998	4
Resume	Executives' Resumes from Chinese Listed Companies	4,761	8
Bank	Banking Loan Data NER	10,000	4
FNED	Domain Event Detection under High Robustness Requirements	10,500	7
DLNER	Discourse-level NER	28,897	9

口语化文本：
哎张姐，我昨天带老爷子去医院复查，医生让做了个X光胸片正侧位检查。您说这检查为啥要连续做三到四次啊？每天早上都得折腾一趟。王主任还建议做个CT断层扫描，说要看清楚肺部纹理和结节病灶。不过隔壁床的李叔说他只做了MRI核磁共振，现在这医疗技术真是五花八门的...
实体列表：X光胸片正侧位检查；CT断层扫描；MRI核磁共振；肺部纹理；结节病灶 领域标签：医疗健康

示例2
原始文本：泗水县政府办今年3月份开始禁止私挖砭石，许多村庄还组织了巡逻队，发现盗挖情况立即阻止。
实体列表：泗水县政府办；巡逻队
生成的口语化文本、实体列表以及领域标签：
口语化文本：
哎老王，你听说没？县里新发的红头文件可严啦！就三月份那会儿，泗水县政府办专门派了巡查组到咱村，说是砭石矿脉现在严禁私挖。昨儿个张大爷家的二小子在后山用金属探测器找矿苗，刚刨两锄头就被护山队逮个正着。要我说啊，这玄武岩层的地质保护确实得抓，上个月邻村就因为过度开采闹出地面塌陷了。对了，环保局还说要给咱配手持式光谱仪，方便检测矿石成分呢。
实体列表：泗水县政府办；巡查组；砭石矿脉；金属探测器；护山队；玄武岩层；地面塌陷；环保局；手持式光谱仪
领域标签：自然资源管理

现在请你根据上述要求并结合示例的回复格式，参考提供的原始文本和实体列表，生成口语化文本、实体列表和领域标签。

原始文本：{raw_text}
实体列表：{entities}

生成的口语化文本、实体列表以及领域标签：

B.3 English Prompt For ContextASR-Dialogue Generation

 # Task
Generate a natural multi-person conversation script in English about a specific movie, along with an entity list, based on provided information.

 # Input Details
1. Movie Title: Title of the film being discussed.
2. Director: Name of the film's director.
3. Cast: Main actors/actresses in the film.
4. Plot Summary: Brief synopsis of the movie's storyline.
5. Number of Participants: Total people in the conversation.

 # Output Requirements
1. Dialogue Script: Casual conversation in screenplay format ("Speaker: Dialogue").
2. Entity List: Proper nouns (titles, names) and film-related terminology from the conversation (semicolon-separated).

 # Key Specifications
1. Participant Names: Use culturally appropriate Western names (e.g., Chris, Emily, Marcus, Rachel) matching participant count.
2. Natural Dialogue: The content of the conversation must be highly colloquial, natural and fluent, in line with the true context of easy discussion of the movie between friends or fans, avoiding any written language or blunt wording, and should be full of life and personal opinions.
3. Movie Content Focus: The core content of the discussion must revolve around the provided movie, engaging in an in-depth analysis. If the provided plot summary is not detailed enough, please reasonably supplement and expand by incorporating information you are aware of regarding the movie (such as a more detailed plot, background, themes, reviews, etc.) to enrich the conversation content, making it more profound and comprehensive. You may include perspectives on the plot, characters, actors' performances, directorial techniques, thematic significance, and other aspects.
4. Entity Integration: Mention at least 3 movie-related names (director/actor/character), with at least 2 names naturally mentioned by different speakers. Please pay special attention that all names mentioned in the dialogue must be in their full form. For example, use "Robin White" instead of just "White".
5. Dialogue Coherence: The entire generated dialogue text should possess a high degree of realism and logical coherence, ensuring that statements between different speakers naturally follow, respond to, and advance the discussion topic, forming an organic and complete dialogue process.
6. Text Normalization: The dialogue text needs to undergo "normalization" processing to simulate speech transcription effects. All numbers (such as years, times, quantities, rankings, dates, currencies, units, etc.) should be converted into their corresponding English words (for example, "nineties" instead of "90s", "eight-thirty" not "8:30", "three hundred dollars" not "300$") instead of using Arabic numerals or symbols.
7. Entity Extraction: Carefully review the generated dialogue text, extract all movie titles, directors, actors, main character names, as well as professional terms related to movie production and film reviews mentioned in the dialogue text, and organize them into a list separated by semicolons ";". Please ensure that all entities in the list accurately appear in the generated dialogue text.
8. Format Rules: It is strictly prohibited to use any structured markup languages (such as Markdown, LaTeX, HTML, XML, JSON, etc.) or emoticons in the generated dialogue text. The dialogue content should only include English words or letters, and common punctuation marks, with each line presented in the format "Speaker: Content". Each speaker in the entire dialogue must have no fewer than three utterances.

 # Language Specific Instructions
The provided movie information is in Chinese (including movie titles, directors' and cast members' full names, and plot summary). When generating English dialogue based on provided Chinese movie information ensure the following:
1. Movie Title: Do not simply translate the Chinese titles. Instead, use the official English titles of the movies. For example, translate "怦然心动" to "Flipped" rather than a literal translation like "Heart pounding" or something.
2. Personal Names: Use the authentic English names of directors, cast members and characters instead of directly transliterating from Chinese.
3. Plot Summary: While translating the plot summary, maintain the original context and nuance without altering the intended meaning. Ensure that cultural references are appropriately adapted for an English-speaking audience.
 # Example
Movie Title: 《花生酱猎鹰》
Director: 泰勒·尼尔森 / 迈克·舒瓦茨
Cast: 希亚·拉博夫 / 达科塔·约翰逊 / 扎克·高察根 / 约翰·浩克斯
Plot Summary: 无亲无故的扎克（扎克·高察根 Zack Gottsagen 饰）是一名唐氏综合症患者，政府不得不将他安置到了一家养老院中，在这里，善良的扎克非常讨人喜欢，与此同时，他亦和在这里工作的名叫爱丽诺（达科塔·约翰逊 Dakota Johnson 饰）的护工之间结下了深厚的友谊。i扎克是一个摔角迷，每天，他都要看一段摔跤比赛的录像，这段录像是由一个化名"咸水乡巴人"的人录制的。扎克的人生理想就是能够前往录像中所记录的摔跤学校。在室友卡尔（布鲁斯·邓恩 Bruce Dern 饰）的建议下，扎克开始尝试"越狱"，并且最终成功，就这样，他踏上了圆梦之旅。
Number of Participants: 3
Generated dialogue script and entity list:
Dialogue Script:
Sarah: Okay, so who else just rewatched The Peanut Butter Falcon this weekend? I need to talk about it!
Jake: Me! That movie hits different every time. Shia LaBeouf as Tyler is so good. His chemistry with Zack Gottsagen—actual magic.
Mia: Right? And Dakota Johnson as Eleanor, the nurse? She's underrated. That scene where she finally lets Zack go chase his wrestling dream—I totally cried when she said, "You're my friend first."
Sarah: Ugh, same. Did you all know Zack Gottsagen is the first actor with Down syndrome to headline a major studio film? That's huge.
Jake: For real? No wonder it felt authentic. Directors Tyler Nilson and Michael Schwartz nailed the balance between humor and heart. Like, the river scenes? When Tyler teaches Zack to crab-catch? Pure joy.
Mia: And don't forget John Hawkes as the creepy fisherman! His whole "saltwater redneck" vibe was low-key terrifying. But Zack's obsession with that wrestling tape? Classic.
Sarah: The way Zack quotes "Salty Water Redneck" like it's scripture… iconic. But the best part? How Tyler's own redemption arc ties into Zack's dream. They both need each other.
Jake: Exactly! It's not just a "road trip" movie. It's about broken people fixing each other. Shia's breakdown scene in the firelight? Oscar-worthy.
Mia: Think they'll ever make a sequel? Zack ruling the wrestling world? Sarah: Nah, the ending is perfect. He finally gets his hero moment in the ring—no CGI,

no cheesy dialogue. Just raw celebration.
Jake: Plus, real-life Zack Gottsagen helped write his lines. You can tell. That "I'm a wrestler, not a criminal" speech? Chills.
Entity List: The Peanut Butter Falcon; Tyler Nilson; Michael Schwartz; Shia LaBeouf; Zack Gottsagen; Dakota Johnson; John Hawkes; Tyler; Eleanor; Salty Water Redneck; Down syndrome; studio film; redemption arc; CGI; Oscar-worthy

Now, please generate dialogue script and entity list based on the above requirements and the provided movie information, referring to the example's reply format.

Movie Title: {movie_name}
Director: {movie_director}
Cast: {movie_actors}
Plot Summary: {movie_plot}
Number of Participants: {person_num}
Generated dialogue script and entity list:

B.4 Chinese Prompt For ContextASR-Dialogue Generation

任务
根据提供的电影元数据信息，生成一段多人参与的、围绕该电影展开讨论的口语化对话文本（剧本格式）。

输入详情
1. 电影名称：要讨论的电影的片名。 2. 电影导演：该电影的导演姓名。 3. 电影演职人员表：参与该电影的主要演员列表。 4. 电影剧情简介：一段简要概括电影核心剧情的文字。 5. 讨论人数：参与本次对话讨论的总人数。

输出详情
1. 对话文本：生成的口语化对话内容，组织为剧本形式，每行格式为"说话人：具体内容"。
2. 实体列表：对话中出现的电影名称、导演、演员、主要角色名称等命名实体以及对话文本中出现的相关专业术语列表（使用分号"；"隔开）。

具体要求
1. 设定说话人：根据提供的讨论人数，请预先设定相应数量的、具有区分度的名字作为对话的参与者（例如：小王、张哥、丽丽等），并在生成的对话文本中使用这些名字作为说话人标识。
2. 口语化风格：对话内容必须高度口语化、自然流畅，符合朋友或影迷之间轻松讨论电影的真实语境，避免任何书面语、正式表达或生硬的措辞，应充满生活气息和个人观点。
3. 聚焦电影内容：对话的核心内容必须围绕提供的电影名称及其剧情展开深入讨论。如果提供的剧情简介不够详细，请结合你已知晓的该电影相关信息（如更详细的剧情、背景、主题、影评等）进行合理补充和扩展，以丰富对话的内容，使其更具深度和广度，可以涉及对剧情、角色、演员表演、导演手法、主题意义等方面的看法。
4. 实体提及要求：在对话过程中，必须提及至少三个和电影相关的人物名字，可以是导演、演员的名字，或剧中的角色名。所有被提及的姓名中，至少有两个必须出现在至少两位不同说话人的对话内容中。请确保这些姓名是自然融入对话而非刻意堆砌。
5. 对话连贯性：生成的整段对话文本应具备高度的真实感和逻辑连贯性，确保不同说话人之间的发言能够自然地承接、回应并推进讨论话题，形成一个有机、完整的对话过程。
6. 文本正则化：对话文本需进行"正则化"处理，模拟语音转录效果。所有出现的数字（如年份、时间、数量、排名、日期、货币、单位等）都应转化为对应的汉字或词语形式（例如，"一九九四年"、"五月一日"、"晚上八点"、"两百块钱"、"十公斤"、"第五名"），而不是使用阿拉伯数字或符号。
7. 实体列表整理：仔细梳理生成的对话文本，提取所有出现的电影名称、导演、演员、主要角色名称以及对话文本中涉及到的相关电影制作、影评等方面的专业名词或术语，组织成列表，用分号"；"隔开。请确保列表中所有的实体都准确无误地出现在生成的对话文本中。
8. 格式要求：严禁在生成的对话文本中使用任何结构化标记语言（如Markdown、LaTeX、HTML、XML、JSON等）或表情符号。对话内容应仅包含汉字、字母和常见的标点符号，每行以"说话人：内容"的格式呈现，整个对话中每个说话人不能少于3次发言。

示例 电影名称：《破·地狱》
电影导演：陈茂贤
电影演职人员表：黄子华 / 许冠文 / 卫诗雅 / 朱康
电影剧情简介：婚礼策划师道生（黄子华 饰）因婚礼市场萧条而债台高筑，被迫改行成为葬礼经纪人。红白二事大相径庭，令道生处处碰壁，但最难一关是要得到喃呒师傅文哥（许冠文 饰）的认可。起初因为理念不合，道生与文哥冲突不断，两人的关系岌岌可危。但数次危难时刻文哥的出手相助，以及亲历文哥与女儿文（卫诗雅 饰）的相处点滴，道生与文哥之间的心结慢慢解开，也逐渐悟到"破地狱"的真正意义。
讨论人数：4
生成的的对话文本和实体列表：
对话文本：
小王：你们看了陈茂贤新片《破·地狱》没？我昨晚上映首日就冲去看了！
丽丽：我哭湿两包纸巾！黄子华和许冠文那场雨中对手戏绝了，许冠文颤着手点烟那个镜头，影帝不愧 是影帝啊。
张哥：最戳我的是卫诗雅演的女儿送饭那段，她蹲在殡仪馆走廊里擦饭盒，明明没台词但看得人心里发酸。
阿斌：听说剧组专门请了二十多位真实喃呒师傅指导，那个破地狱仪式的特写镜头，香火在镜头前飘散的效果太震撼了。
小王：黄子华这次转型真成功，从油嘴滑舌的婚庆策划到满脸沧桑的葬礼经纪人，他给文哥下跪那场戏，我鸡皮疙瘩都起来了。
丽丽：你们注意到道生办公室那面照片墙没？从婚庆合照慢慢变成葬礼白花，这个细节太扎心了。
张哥：许冠文和卫诗雅的父女线才叫虐，文哥每次欲言又止的表情，把中国式父亲那种别扭的关心演得入木三分。
阿斌：我查资料说"破地狱"本来是要打破亡魂执念，结果道生自己才是那个被困在世俗里的游魂，这隐喻绝了。
小王：结尾烧纸船那幕，火星子飘到天上变成星星，这镜头语言比直接煽情高级多了。
丽丽：不过朱康演的讨债佬是不是太脸谱化了？每次出场都扯着嗓子吼，感觉可以再收着点演。
张哥：我倒觉得这种夸张反而有黑色幽默，特别是他追到殡仪馆被纸人吓到那段，全场都在爆笑。
阿斌：说到配乐，唢呐混电子乐那段葬礼进行曲你们记得吗？传统和现代碰撞得恰到好处。
小王：听说导演为拍火盆反光效果，硬是等了三天的阴雨天，最后呈现的冷色调画面确实有幽冥感。
丽丽：我二刷发现个细节，道生西装口袋永远别着婚庆胸花，直到最后才换成白菊，这角色弧光太完整了！
实体列表：
《破·地狱》；陈茂贤；黄子华；许冠文；卫诗雅；朱康；道生；文哥；葬礼经纪人；喃呒师傅；破地狱仪式；殡仪馆；黑色幽默；唢呐；电子乐；镜头语言；冷色调；角色弧光

现在请你根据上述要求并结合提供的电影信息，参考示例的回复格式，生成对话文本和实体列表。

电影名称：{movie_name}
电影导演：{movie_director}
电影演职人员表：{movie_actors}
电影剧情简介：{movie_plot}
讨论人数：{person_num}
生成的的对话文本和实体列表：

C Detailed Results on Open-Source ASR Benchmarks

To thoroughly demonstrate that existing ASR benchmarks fail to unveil the improvements brought by LLM to speech recognition tasks, we select several representative conventional ASR systems and LLM-based ASR systems and test them on 21 English and 54 Mandarin existing open-source benchmarks. Table 4 and Table 5 show the detailed results on English and Mandarin benchmarks, respectively.

D Prompt used in ContextASR-Bench evaluation

We conduct a comprehensive evaluation and analysis in Section 3.2 to highlight our proposed ContextASR-Bench in assessing how LLMs' strong world knowledge and context learning capabilities enhance contextual speech recognition. Table 6 exhibits user prompt configurations for LALMs under the three context evaluation settings.

Table 4: The WER (%) results of Conventional ASR and Large Audio Languages Models on existing open-source English ASR benchmarks. The "Overall" results represent the average WER of each model on test speeches across all benchmarks.

Dataset	Sensevoice-small	Fireredasr-aed	Whisper-Large-turbo	Whisper-Large-v3	Fireredasr-llm	Qwen2.5omni	Kimi-audio
COMMON_VOICE_V17.0_EN_DEV	13.13	13.88	9.05	8.15	10.94	5.93	6.13
COMMON_VOICE_V17.0_EN_TEST	14.98	17.55	12.04	10.56	15.00	7.72	8.35
FLEURS_EN-US_DEV	8.70	8.09	4.97	4.63	4.88	4.20	5.20
FLEURS_EN-US_TEST	7.83	7.70	4.84	4.59	4.65	3.81	4.63
GIGASPEECH_V1.0.0_DEV	11.61	10.11	11.11	11.45	9.55	11.06	10.44
GIGASPEECH_V1.0.0_TEST	11.73	10.17	10.55	10.64	9.66	11.02	10.06
LIBRISPEECH_DEV_CLEAN	3.49	1.82	2.25	2.24	1.58	1.55	1.60
LIBRISPEECH_DEV_OTHER	6.99	4.22	4.25	3.96	3.06	3.23	2.82
LIBRISPEECH_TEST_CLEAN	3.26	1.94	2.37	2.15	1.65	1.74	1.58
LIBRISPEECH_TEST_OTHER	7.30	4.39	4.36	3.96	3.65	3.47	2.93
MLS_EN_TEST	10.22	7.01	5.39	5.21	5.22	5.45	4.82
SLIDE_SPEECH_DEV	10.44	7.67	9.17	9.53	7.54	8.42	8.18
SLIDE_SPEECH_TEST	11.23	8.18	10.81	10.75	8.11	9.81	8.99
SPGISPEECH_DEV	3.49	5.37	3.33	3.46	4.70	2.25	4.00
SPGISPEECH_TEST	3.50	5.28	3.27	3.36	4.60	2.27	3.91
TEDLIUM_RELEASE3_LEGACY_DEV	4.53	4.75	4.40	4.15	4.11	3.68	3.44
TEDLIUM_RELEASE3_LEGACY_TEST	4.16	3.96	4.27	4.60	3.84	3.93	3.36
VOXPOPULI_V1.0_EN_ACCENTED_TEST	14.16	14.25	18.90	18.77	13.96	23.71	16.52
VOXPOPULI_V1.0_EN_DEV	10.82	10.73	9.99	9.72	10.32	6.70	8.90
VOXPOPULI_V1.0_EN_TEST	10.45	10.61	10.67	9.14	9.96	6.51	8.91
Overall	7.12	7.65	6.50	6.44	6.81	5.81	6.15

Table 5: The WER (%) results of Conventional ASR and Large Audio Languages Models on existing open-source Mandarin ASR benchmarks. The "Overall" results represent the average WER of each model on test speeches across all benchmarks.

Dataset	Sensevoice-Small	Paraformer-Large	FireredASR-AED-L	Whisper-Large-turbo	Whisper-Large-v3	Dolphin-Small	Dolphin-Base	FireredASR-LLM-L	Qwen2.5-Omni	Kimi-Audio
AISHELL1_TEST	3.01	1.93	0.55	6.10	5.48	3.33	4.47	0.73	1.62	0.76
AISHELL2_ANDROID_TEST	3.94	3.08	2.76	5.81	5.21	4.74	5.96	2.50	2.76	2.63
AISHELL2_IOS_TEST	3.81	2.84	2.52	5.55	4.91	4.42	5.58	2.16	2.59	2.84
AISHELL2_MIC_TEST	3.88	3.01	2.81	5.51	5.07	4.78	6.21	2.49	2.63	2.76
AISHELL4_TEST	16.59	17.13	11.79	31.78	29.10	20.15	21.90	12.06	19.26	20.00
ALIMEETING_EVAL_FAR	24.21	21.84	14.22	38.02	36.21	30.68	31.97	14.87	26.81	26.06
ALIMEETING_EVAL_NEAR	5.55	5.15	3.34	14.23	12.19	6.28	7.49	3.97	5.56	6.98
ALIMEETING_TEST_FAR	25.42	23.12	15.44	40.05	37.92	32.61	35.08	16.35	29.86	29.30
ALIMEETING_TEST_NEAR	7.05	6.47	4.09	16.05	15.55	8.38	9.23	4.89	6.87	8.35
ASRU_TEST	8.12	5.34	6.60	10.51	9.70	9.20	11.74	5.71	8.39	7.21
COMMON_VOICE_V17.0_ZH_DEV	13.47	12.95	7.15	17.49	16.66	17.95	16.78	7.20	8.38	9.45
COMMON_VOICE_V17.0_ZH_TEST	10.57	10.24	3.39	14.42	12.84	11.53	14.43	3.51	5.06	6.06
FLEURS_CMN_DEV	3.56	3.34	3.20	4.13	3.53	4.07	5.99	2.41	2.31	2.28
FLEURS_CMN_TEST	4.16	3.80	3.64	4.70	4.08	4.48	6.31	2.54	2.59	2.52
KESPEECH_DEV	8.43	9.53	3.82	24.34	21.03	8.61	14.33	3.17	5.58	4.82
KESPEECH_TEST	10.15	11.37	4.53	34.10	29.18	10.68	15.02	3.60	6.46	5.24
MAGICDATA_CONVERSATION_DEV	8.08	7.81	4.62	18.58	16.46	9.54	11.72	5.10	7.02	25.69
MAGICDATA_CONVERSATION_TEST	10.70	10.56	6.36	22.41	19.73	12.09	14.67	6.92	9.45	36.48
MAGICDATA_READ_DEV	4.36	4.12	0.79	10.57	9.10	5.09	6.84	1.41	2.71	1.69
MAGICDATA_READ_TEST	4.02	4.00	0.92	9.01	8.03	4.29	5.66	1.46	2.71	1.73
SEAME_DEV_MAN	27.09	33.11	31.21	44.78	36.27	37.52	48.98	31.14	31.95	35.12
SEAME_DEV_SEG	39.22	53.73	50.98	91.22	77.66	78.76	131.93	51.75	53.80	54.42
SPEECHIO_ASR_ALL	3.64	2.90	2.82	8.43	7.64	4.89	6.64	2.63	3.20	2.77
SPEECHIO_ASR_ZH00000	2.82	2.58	2.28	9.49	8.92	3.30	4.05	2.30	2.64	2.36
SPEECHIO_ASR_ZH00001	1.04	0.61	0.79	2.45	2.10	1.56	2.32	0.59	0.70	0.52
SPEECHIO_ASR_ZH00002	4.44	3.51	3.00	8.05	7.37	5.02	6.75	2.88	3.50	3.71
SPEECHIO_ASR_ZH00003	2.45	1.19	1.13	5.53	4.66	3.19	5.44	0.95	1.00	0.94
SPEECHIO_ASR_ZH00004	2.22	1.77	1.57	4.00	3.73	2.75	3.58	1.59	2.09	1.61
SPEECHIO_ASR_ZH00005	2.79	2.16	2.24	8.27	8.36	3.72	4.70	2.12	2.62	1.91
SPEECHIO_ASR_ZH00006	6.33	5.26	4.81	16.28	13.65	7.39	9.47	4.79	6.26	5.46
SPEECHIO_ASR_ZH00007	6.71	4.95	3.67	14.49	13.19	10.23	12.65	3.78	7.44	5.42
SPEECHIO_ASR_ZH00008	5.33	4.34	4.10	15.71	14.16	8.50	12.94	4.00	6.65	5.08
SPEECHIO_ASR_ZH00009	4.06	3.41	3.54	7.94	7.28	4.98	6.41	3.31	3.67	3.38
SPEECHIO_ASR_ZH00010	3.87	3.43	3.36	8.79	8.49	4.32	5.26	3.31	3.50	3.53
SPEECHIO_ASR_ZH00011	2.02	1.51	1.37	5.93	5.13	3.18	4.47	1.40	1.56	1.33
SPEECHIO_ASR_ZH00012	3.64	3.04	2.27	10.44	8.45	4.49	5.62	2.06	3.24	2.30
SPEECHIO_ASR_ZH00013	4.17	3.53	4.28	8.16	7.51	5.95	8.74	4.00	3.79	4.08
SPEECHIO_ASR_ZH00014	5.08	4.21	3.53	10.23	8.67	8.41	12.75	3.55	4.20	3.77
SPEECHIO_ASR_ZH00015	7.33	5.21	8.16	16.02	14.72	12.09	15.96	7.00	6.47	5.73
SPEECHIO_ASR_ZH00016	6.83	5.43	5.49	13.49	12.56	9.14	11.66	5.22	5.64	5.15
SPEECHIO_ASR_ZH00017	3.75	2.76	2.65	9.39	7.89	5.53	7.30	2.47	2.94	2.56
SPEECHIO_ASR_ZH00018	3.39	3.14	2.54	5.99	5.74	4.25	6.28	2.44	3.45	3.04
SPEECHIO_ASR_ZH00019	4.32	3.84	3.43	11.95	11.45	7.11	10.31	3.31	4.10	3.30
SPEECHIO_ASR_ZH00020	2.59	1.32	1.63	5.90	5.07	3.76	5.90	1.34	1.58	1.34
SPEECHIO_ASR_ZH00021	3.73	3.10	2.81	7.64	7.43	5.23	7.29	2.72	3.26	2.65
SPEECHIO_ASR_ZH00022	5.87	5.07	4.12	9.92	8.97	7.02	9.45	3.52	4.63	3.39
SPEECHIO_ASR_ZH00023	3.26	2.80	2.48	6.88	6.39	4.45	6.34	2.18	2.59	2.78
SPEECHIO_ASR_ZH00024	6.43	4.97	4.95	18.36	15.95	10.08	12.53	4.55	5.90	4.99
SPEECHIO_ASR_ZH00025	5.05	4.49	3.87	11.06	10.45	6.42	10.83	3.65	4.53	4.37
SPEECHIO_ASR_ZH00026	4.79	4.14	4.32	7.25	6.88	5.51	7.43	3.99	4.62	3.57
THCHS-30_DEV	4.72	3.76	0.09	7.50	6.35	5.05	7.54	0.32	2.71	1.10
THCHS-30_TEST	5.18	3.98	0.27	7.62	6.83	5.67	7.74	0.56	3.07	1.36
WENETSPEECH_DEV	3.49	3.14	3.21	9.59	8.86	7.53	8.39	3.23	4.72	3.11
WENETSPEECH_TEST_MEETING	7.34	6.98	4.76	19.03	18.94	7.83	10.11	4.63	7.64	6.23
WENETSPEECH_TEST_NET	7.13	6.63	4.85	11.55	9.90	9.30	11.74	4.60	5.97	6.44
Overall	8.46	8.18	5.66	16.88	15.19	10.86	13.52	5.68	8.03	9.84

Table 6: User prompts used in ContextASR-Bench evaluation for Large Audio Language Models (LALMs). All LALMs are evaluated on ContextASR-Bench with different user prompts under three context evaluation settings. "Context" refers to context evaluation settings, where "/", "Coarse", and "Fine" represent Contextless, Coarse-grained Context, and Fine-grained Context settings. ContextASR-Bench includes two distinct test sets: ContextASR-Speech (denoted "Speech") and ContextASR-Dialogue (denoted as "Dialogue"), while "Both" indicates both sets. "<domain label>" and "<movie name>" indicate Coarse-grained Context for the ContextASR-Speech and ContextASR-Dialogue sets, respectively. Both sets use "<entity list>" as Fine-grained Context.

Models	Language	Context	Test Set	Prompt
Qwen2-Audio	EN / ZH	/	Both	Detect the language and recognize the speech:
		Coarse	Speech	This speech belongs to the <domain label> field. Detect the language and recognize the speech:
			Dialogue	This speech is a dialogue about the movie <movie title> . Detect the language and recognize the speech:
		Fine	Speech	This speech belongs to the <domain label> field and may contains the following words or phrases: <entity_list>. Detect the language and recognize the speech:
			Dialogue	This speech is a dialogue about the movie <movie title> and may contains the following words or phrases: <entity_list>. Detect the language and recognize the speech:
Kimi-Audio	EN / ZH	/	Both	Please transcribe the following audio:
		Coarse	Speech	The following audio belongs to the <domain label> field. Please transcribe the following audio:
			Dialogue	The following audio is a dialogue about the movie <movie title> . Please transcribe the following audio:
		Fine	Speech	The following audio belongs to the <domain label> field and may contains the following words or phrases: <entity_list>. Please transcribe the following audio:
			Dialogue	The following audio is a dialogue about the movie <movie title> and may contains the following words or phrases: <entity_list>. Please transcribe the following audio:
Baichuan-Audio **Baichuan-Omni-1.5**	EN / ZH	/	Both	将语音转录为文本:
		Coarse	Speech	这段语音属于 <domain label> 领域。 将语音转录为文本:
			Dialogue	该语音是一段讨论电影 <movie name> 的对话。 将语音转录为文本:
		Fine	Speech	这段语音属于 <domain label> 领域，并且可能包含以下词或短语：<entity_list>。将语音转录为文本:
			Dialogue	该语音是一段讨论电影 <movie name> 的对话，并且可能包含以下词或短语：<entity_list>。将语音转录为文本:
Qwen2.5-Omni	EN	/	Both	Transcribe the English audio into text, ensuring all punctuation marks are included.
		Coarse	Speech	This audio belongs to the <domain label> field. Transcribe the English audio into text, ensuring all punctuation marks are included.
			Dialogue	This audio is a dialogue about the movie <movie title> . Transcribe the English audio into text, ensuring all punctuation marks are included.
		Fine	Speech	This audio belongs to the <domain label> field and may contains the following words or phrases: <entity_list>. Transcribe the English audio into text, ensuring all punctuation marks are included.
			Dialogue	ThThis audio is a dialogue about the movie <movie title> and may contains the following words or phrases: <entity_list>. Transcribe the English audio into text, ensuring all punctuation marks are included.
	ZH	/	Both	请将这段汉语语音转换为带有标点符号的文本。
		Coarse	Speech	这段语音属于 <domain label> 领域。 请将这段汉语语音转换为带有标点符号的文本。
			Dialogue	该语音是一段讨论电影 <movie name> 的对话。 请将这段汉语语音转换为带有标点符号的文本。
		Fine	Speech	这段语音属于 <domain label> 领域，并且可能包含以下词或短语：<entity_list>。请将这段汉语语音转换为带有标点符号的文本。
			Dialogue	该语音是一段讨论电影 <movie name> 的对话，并且可能包含以下词或短语：<entity_list>。请将这段汉语语音转换为带有标点符号的文本。