PARROT: SEAMLESS SPOKEN DIALOGUE INTERACTION WITH DOUBLE-CHANNEL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advancements in large language models (LLMs) have demonstrated significant potential in enhancing real-time spoken interactions. Presently, open-source methodologies predominantly depend on intermediate generative text-based transcriptions to manage real-time spoken dialogues. However, these techniques often struggle with providing seamless interactions that involve real-time streaming audio inputs. In this research, we unveil an innovative spoken dialogue language model, **Parrot**, distinguished by its unique pre-training and supervised fine-tuning (SFT) pipeline. This pipeline deviates from conventional methodologies by utilizing both single-channel audio data and double-channel spoken dialogue data to train the textless speech language model. During pre-training, we transform single-channel audio input into a sequence of discrete tokens, thereby instructing the LLM to identify audio tokens via next-token predictions. In the SFT phase, we pioneer a novel approach to double-channel generative spoken dialogue language modeling with a unique "next-token-pair prediction" objective, facilitating the LLM's comprehension of natural human conversations. Our pipeline equips LLM to produce spoken interactions that are more natural and fluid than those generated by baseline approaches, as substantiated by thorough evaluations[1].

1 INTRODUCTION

The advent of large language models (LLMs), particularly the GPT series (Patel et al., 2023; OpenAI, 2023; 2024), has profoundly transformed the field of artificial intelligence. These powerful language models attain their capabilities through pretraining on extensive text corpora using decoder-only transformer architectures, guided by an autoregressive next-token prediction objective function. Recently, there has been an increasing interest in integrating the LLMs with other modalities, such as images (Radford et al., 2021; Li et al., 2022; 2023; Liu et al., 2023b), audio (Zhang et al., 2023a; 2024a; Hassid et al., 2023), protein sequences (Lin et al., 2022; Madani et al., 2023) and etc . Among these modalities, audio or speech data holds particular importance as it enables LLMs to engage in real-time voice interactions with humans. The recently unveiled GPT-4o model (OpenAI, 2024) exhibits a remarkable proficiency in managing real-time interactions with users in conversational contexts. Throughout the demo presentation, it was able to generate authentic emotional responses and engage users with swift reactions. These functionalities, however, introduce additional challenges, as the model must thoroughly interpret the distinct audio information within human speech while conducting inference with minimal delay.

Presently, the academic community primarily utilizes open-sourced models (Zhang et al., 2023a; Xie & Wu, 2024; Rubenstein et al., 2023; Huang et al., 2024; Wang et al., 2023a; Nachmani et al., 2024; Wang et al., 2023b) following a cascading approach. This method heavily depends on an intermediate text generation step and generally consists of three stages: automatic-speech-recognition (ASR), text-based question answering (Text-QA), and text-to-speech (TTS) synthesis. While this approach is reliable due to the incorporation of powerful text-based LLMs, it does present three significant drawbacks: (1) **Audio Information Loss:** Audio signals, unlike text, include additional

[1]Demo and code can be found at `https://anonymous.4open.science/r/Parrot`.

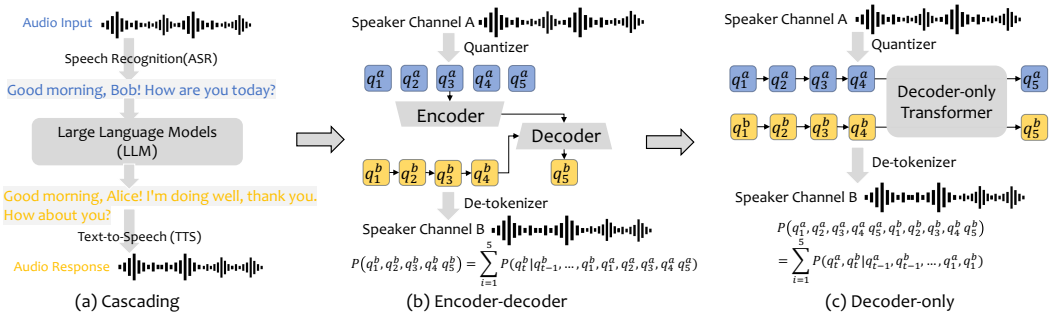


Figure 1: (a) The cascading approach depends on the intermediate text-based response generation translated by ASR and TTS; (b) The encoder-decoder spoken dialogue language modeling encode one of the speaker's audio sequence $Q^a = (q_1^a, q_2^a, ..., q_T^a)$ as condition information to decode another speaker sequence Q^b following the probability distribution $P(Q^b) = \sum_{i=1}^{T} P(q_t^b | q_{t-1}^b, ..., q_1^b, Q^a)$; (c) Our novel decoder-only spoken dialogue language modeling follows the newly proposed next-token-pair prediction paradigm such that $P(Q^a, Q^b) = \sum_{i=1}^{T} P(q_t^a, q_t^b | q_{t-1}^b, ..., q_1^b, q_{t-1}^a, ..., q_1^a)$.

human responses such as laughter, interruptions, pauses, and repetitions, reflecting the speaker's communication style and emotions. The conversion of audio signals to text could potentially result in the loss of this crucial information. (2) **Error Propagation:** The cascading approach consists of three sequential stages. If the initial ASR translation is inaccurate, the subsequent stages will operate on incorrect intermediate data representations. (3) **Real-time Processing Challenges:** In real-world applications, such as the GPT-4o presentation, spoken dialogues require immediate processing. However, incorporating text translation steps inevitably results in a slower process and adds extra latency during inference. Many recently introduced speech LLMs are striving to mitigate these issues. However, they either depend on text generation or remain confined to basic question-answer functions. We will delve into detailed discussions about these approaches in the subsequent related work section and the appendix, given the rapid growth of this research field. Therefore, the aforementioned limitations of cascading approaches highlight the necessity of developing speech-to-speech models capable of managing spoken conversations without the need for text translations.

In this study, we present a novel pre-training and supervised fine-tuning (SFT) pipeline to develop a robust model, referred to as **Parrot**, specifically designed for spoken dialogue language modeling. The pre-training phase begins with the conversion of continuous audio inputs into a sequence of tokens, a process made possible by training a vector-quantized autoencoder (VQVAE) (van den Oord et al., 2017) to reconstruct these audio signals. We then leverage pretrained LLMs as a foundation for continuous learning on *single-channel* audio sequences, with the goal of next-token prediction. This is accomplished by integrating the learned audio tokens into the original text vocabulary. This pretraining stage aids LLMs in capturing the primary latent distribution of audio token sequences. In the subsequent stage, we utilize *double-channel* audio data for SFT. The key advantage is enabling LLMs to directly comprehend how humans engage in natural dialogues. Unlike existing approaches, we introduce a novel "**next-token-pair prediction**" paradigm to model the double-channel spoken dialogue generation using the decoder-only transformer. The comparison between our proposed method and existing techniques are illustrated in Figure 1. We carry out extensive experiments to validate the superiority of our innovative approach. Specifically, **Parrot** consistently outperforms strong baseline methods by 150% and 200% in average in terms of the reflective pause and interruption response accuracy respectively. Additionally, it achieves a low latency of 300ms. In summary, our contributions to the field are as follows:

1) We present a spoken dialogue language model, **Parrot**, featuring an innovative pre-training and SFT pipeline. This novel approach eliminates the need for intermediate text conversions, thereby facilitating more fluid and natural voice interactions with human users at reduced latency.
2) We propose a new paradigm for double-channel spoken language modeling, called next-token-pair prediction, which holds the potential to be readily generalized for autoregressive modeling of multi-channel audio sequence inputs in future explorations.

3) We provide an extensive evaluation of spoken dialogue language models, encompassing several key aspects and metrics for assessing the quality and speed of spoken interactions.

2 RELATED WORKS

Autoregressive Generative Models. The autoregressive generative modeling has achieved remarkable success in natural language processing, giving rise to a variety of powerful LLMs (Sutskever et al., 2014; OpenAI, 2024; 2023; Patel et al., 2023). Inspired by these LLMs, numerous studies have examined the application of autoregressive modeling in other domains, such as images (van den Oord et al., 2017; Esser et al., 2021; Li et al., 2024; Tian et al., 2024; Lee et al., 2022; Chang et al., 2022), graphs (You et al., 2018), videos (Weissenborn et al., 2020), molecules (Shi et al., 2020; Schwaller et al., 2019) and protein sequences (Madani et al., 2023; Lin et al., 2022). The fundamental concept of autoregressive modeling focuses on iteratively generating the entire segment from the intermediate portion, which is particularly well-suited for the audio generation.

Multi-modal LLMs. Multimodal Large Language Models (MM-LLMs) strive to incorporate knowledge from diverse modalities. A key category of MM-LLMs concentrates on developing connectors (Li et al., 2022; 2023; Liu et al., 2023b; Alayrac et al., 2022) that identify knowledge alignment across various modalities. An alternative strategy (Team, 2024; Zhou et al., 2024; Xie et al., 2024) merges all modalities into a cohesive sequence of tokens and utilizes LLMs to sequentially generate them using modified attention masks. These methods (Wu et al., 2024; Su et al., 2023; Fu et al., 2024) even integrate audio as an input modality, and by simply combining text and audio through MM-LLM techniques, they can address one-direction conditional generation tasks such as speech-to-text translation (e.g., ASR and spoken language understanding) (Radford et al., 2023; Zhang et al., 2023b; Deshmukh et al., 2023; Arora et al., 2023; Tang et al., 2024; Chu et al., 2024; Zhou et al., 2023; Ravanelli et al., 2021; Gao et al., 2023) and text-to-speech translation (e.g., TTS) (Elizalde et al., 2023; Liu et al., 2023a; Huang et al., 2023; Nachmani et al., 2023; Yang et al., 2023; Kreuk et al., 2023; Borsos et al., 2023; Copet et al., 2023; Chen et al., 2024; Anastassiou et al., 2024; Jiang et al., 2023b; Kong et al., 2021; Shen et al., 2024; Casanova et al., 2022; Siuzdak, 2024; Yang et al., 2024; Kharitonov et al., 2023; Le et al., 2023). However, these methods are limited to handling multi-turn multi-modal QA tasks (where the model produces an answer only after the question is completed, as signaled by pressing the input button, for instance) and thereby struggle with real-time voice interaction tasks, which is the primary focus of our work.

Generative Spoken Language Modeling. The core concept of our approach relies on the pretraining of robust speech foundation models, with language model learning (LLM) serving as a crucial component, to enable rapid adaptation to a broad spectrum of downstream speech tasks. Much of the prior research has utilized the encoder-decoder architecture to enhance pre-training (Borsos et al., 2023; Lakhotia et al., 2021; Kharitonov et al., 2022; Polyak et al., 2021; Chen et al., 2023; 2022; Hsu et al., 2021; Zeghidour et al., 2022; Défossez et al., 2023; Agostinelli et al., 2023; Ao et al., 2022; Tang et al., 2022; Wu et al., 2023). However, this architecture proves inadequate for handling real-time speech interactions with streaming audio inputs, as it requires the encoder to process the entire input simultaneously. In more recent studies, the decoder-only transformer (Maiti et al., 2024; Zhang et al., 2024a; Hassid et al., 2023; Nguyen et al., 2024; Fathullah et al., 2024; Shen et al., 2023; Zhang et al., 2024b; Das et al., 2024) has been employed to model the audio sequence. This approach capitalizes on the potent language capabilities of LLMs while also facilitating the processing of streaming inputs. Motivated by the advent of GPT-4o, newly developed models aim to endow LLMs with speech conversation capabilities (Ma et al., 2024; Zhang et al., 2023a; Xie & Wu, 2024; Rubenstein et al., 2023; Huang et al., 2024; Wang et al., 2023a; Nachmani et al., 2024; Wang et al., 2023b; Défossez et al., 2024). However, these models either rely on text transcriptions or adhere to the aforementioned MM-LLM methods, lacking the ability for natural turn-taking. In stark contrast, our work leverages double-channel spoken dialogue data to directly instruct LLMs in human conversations. A significant contribution in the realm of spoken dialogue language modeling is dGLSM (Nguyen et al., 2023), but it remains confined to the era of using the encoder-decoder architecture. In our research, we elevate the architecture to the most recent decoder-only transformer.

3 PARROT: TRAINING AND INFERENCE PIPELINE

Our **Parrot** comprises two essential steps. The first involves pretraining the LLM on single-channel audio token sequences using the traditional "next-token prediction" objective. The second step fine-tunes the LLM on double-channel audio token sequences, employing the innovative "next-token-pair prediction" paradigm. The rationale behind this strategy stems from the fundamental observation that the single-channel audio data can be sourced from the vast amount of open-source data available on the web. However, the primary limitation of single-audio data is its lack of speaker identity information and the overlapping regions between different speakers can be misleading. On the other hand, double-channel spoken dialogue data encapsulates crucial turn-taking events with distinct speaker channels, and any overlapping event can be easily discerned. Nevertheless, the double-channel data necessitates specific pre-processing techniques to segregate the mixed information from the single-channel data. Therefore, it is a naturally inspired strategy to use the large-scale single-channel audio data for pretraining and the moderate-scale double-channel dialogue data for SFT.

3.1 AUDIO TOKENIZATION AND SINGLE-CHANNEL AUDIO PRETRAINING

A single-channel audio is a continuous input sequence $\mathbf{x} \in \mathbb{R}^T$ with time length T. Owing to the high sampling rate of continuous audio signals, it is essential to employ an audio tokenizer, which extracts valuable features for the purpose of compressing the information. The audio quantizer $\mathcal{Q}$ projects the audio sequence $\mathbf{x}$ into a set of discrete tokens $Q = (q_1, ..., q_{T'}) = \mathcal{Q}(\mathbf{x})$ ($T' \ll T$), where each token q_t is an integer index from the vocabulary $q_t \in [V]$ where the vocabulary size is V. We train the audio tokenizer $\mathcal{Q}$ following the VQ-VAE (van den Oord et al., 2017) framework. In contrast to certain prior studies, we directly train the tokenizer on the raw audio signals $\mathbf{x}$, rather than transforming $\mathbf{x}$ into a mel-spectrogram first. We primarily adopt the training strategy presented in SoundStream (Zeghidour et al., 2022), and provide a brief overview of its underlying mechanism.

Specifically, audio inputs $\mathbf{x}$ is fed into an encoder $\mathcal{E}$ to derive down-sampled latent features $\mathbf{f} \in \mathbb{R}^{\frac{T}{r} \times D}$ such that $\mathbf{f} = \mathcal{E}(\mathbf{x})$ with the down-sampling rate r and the latent dimension D. This is achieved by the CNN (Krizhevsky et al., 2012) architecture, which can capture the local dependency of $\mathbf{x}$. Then the quantizer $\mathcal{Q}$ converts the latent feature $\mathbf{f}$ to discrete tokens $\mathbf{q} \in \mathbb{R}^{\frac{T}{r}}$ such that $\mathbf{q} = \mathcal{Q}(\mathbf{f})$ where each entry q_i is a quantized integer index. Each latent feature $\mathbf{f}_i$ for time frame i is mapped to the code index q_i of its nearest embedding vector in the Euclidean sense:

$$q_i = \underset{v \in [V]}{\arg\min} \|\mathbf{z}_v - \mathbf{f}_i\|_2, \tag{1}$$

where $\mathbf{z}_i$ denotes the ith embedding vector of the learnable codebook $\mathbf{z} \in \mathbb{R}^{V \times D}$ containing $|V|$ vectors. Then the reconstructed audio signals $\hat{\mathbf{x}}$ are obtained through the decoder $\mathcal{G}$ such that $\hat{\mathbf{x}} = \mathcal{G}(\mathbf{z}_q)$ where $\mathbf{z}_q \in \mathbb{R}^{\frac{T}{r} \times D}$ denotes the codebook embedding vectors of the latent feature $\mathbf{f}$ indexed by $\mathbf{q}$. This autoencoder is trained by both the reconstruction loss and discriminator loss through straight-through estimators with stop-gradient operations. We direct readers to (Zeghidour et al., 2022) for a comprehensive description of the architectures and algorithms involved.

After converting the input audio signals into the sequence of audio tokens Q, we subsequently supplement these audio tokens into the original LLM's text token vocabulary. Following this, we train the LLMs on the sequence Q using the standard autoregressive approach with the next-token prediction paradigm:

$$p(q_1, q_2, ..., q_{T'}) = \prod_{t=1}^{T'} p(q_t | q_{t-1}, .., q_2, q_1). \tag{2}$$

The next-token prediction loss is calculated by summing the cross-entropy loss, which measures the classification over codebook embedding indices at each time step. Instead of building the audio language model from scratch, we employ Llama 3 as the initial LLM, augmenting its vocabulary with additional audio tokens. While LLMs, after the aforementioned pretraining, can learn the basic audio token distribution, relying solely on single-channel audio data is inadequate for LLMs to effectively comprehend the subtleties of human communication and generate smooth, natural responses. Consequently, we continue to train the speech LLM to learn human speech conversations by utilizing double-channel spoken dialogue data.

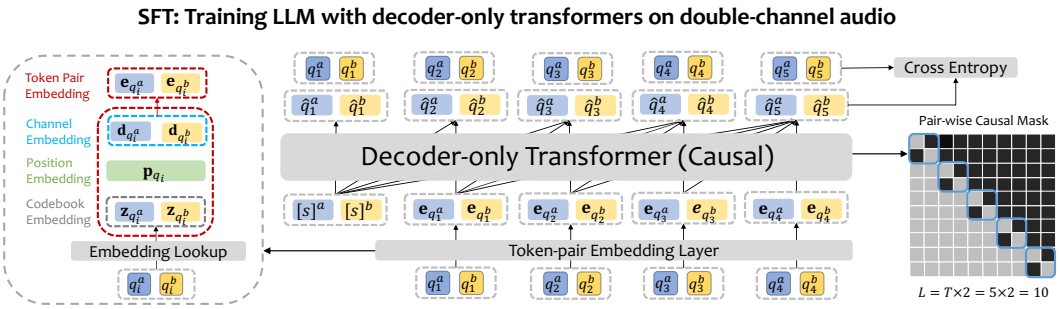


Figure 2: The illustration of the SFT learning mechanism of **Parrot** on the double-channel spoken dialogue data. The novel architecture consists of two important modules. The first module is the embedding layer for obtaining the token-pair embedding; The second module is the decoder-only transformer with a *pair-wise causal masking* attention for next-token-pair prediction. $[s]^a$ and $[s]^b$ denote the special start tokens of channel a and channel b respectively.

3.2 Supervised Fine-tuning with Double-channel Audio

The double-channel audio input comprises a pair of time-aligned single-channel audio inputs, denoted as $(\mathbf{x}^a, \mathbf{x}^b)$, where each channel corresponds to a specific speaker. A fresh challenge arises in the generative modeling of double-channel audio sequences using the decoder-only transformer architecture of LLMs. To address this issue, we propose a novel generative learning paradigm called *next-token-pair prediction*. The key idea here is to generate a sequence of time-aligned token pairs, rather than a single token, in an autoregressive fashion. In contrast to the conventional next-token prediction, our objective is more suitable to the generative modeling of an interpolated dialogue sequence which contain two separate channel identities. Specifically, we begin by discretizing both channels into time-aligned sequences with quantized audio tokens, denoted as $(Q^a = (q_1^a, q_2^a ..., q_T^a), Q^b = (q_1^b, q_2^b ..., q_T^b))$. To accommodate the input sequence structure within the decoder-only transformer architecture, we reorganize both sequences into a single interpolated dialogue sequence, represented as $Q^{\text{input}} = \{q_1^a, q_1^b, q_2^a, q_2^b, ..., q_T^a, q_T^b\}$. Subsequently, we model the probability distribution that generates the next token pair (q_t^a, q_t^b) at next time step t conditioned on the previously generated token pairs from step 1 to $t-1$:

$$p(q_1^a, q_1^b, q_2^a, q_2^b ..., q_T^a, q_T^b) = \prod_{t=1}^{T} p(q_t^a, q_t^b | q_{t-1}^a, q_{t-1}^b, ..., q_2^a, q_2^b, q_1^a, q_1^b). \tag{3}$$

Then we decompose the token pair conditional generating distribution $p(q_t^a, q_t^b | q_{t-1}^a, q_{t-1}^b, ..., q_1^a, q_1^b)$ by assuming the conditional independence between q_t^a and q_t^b:

$$p(q_t^a, q_t^b | q_{t-1}^a, q_{t-1}^b, ..., q_1^a, q_1^b) = p(q_t^a | q_{t-1}^a, q_{t-1}^b, ..., q_1^a, q_1^b) p(q_t^b | q_{t-1}^a, q_{t-1}^b, ..., q_1^a, q_1^b). \tag{4}$$

We illustrate this conditional independence and the dialogue distribution modeling in Figure 1. The probability distribution in Eq.3 and Eq.4 adheres to a fundamental inductive bias that *a person's speech is influenced by both his own previous statements and what he has heard in the past*. To adapt to the generative modeling of the newly arranged dialogue sequence Q^{input}, we need to modify the embedding layer and the attention masking mechanism accordingly. Our novel token-pair embedding layer consists of three important embeddings in total, which are codebook embedding $\mathbf{z}$, position embedding $\mathbf{p}$ and channel embedding $\mathbf{d}$. Specifically, for each token pair q_t^a, q_t^b:

$$\mathbf{z}_{q_t^a}, \mathbf{z}_{q_t^b} = \text{lookup}(\mathbf{z}, q_t^a, q_t^b), \quad \mathbf{p}_{q_t^a} = \mathbf{p}_{q_t^b}, \quad \mathbf{d}_{q_t^a}, \mathbf{d}_{q_t^b} = \text{one-hot-embedding}(\mathbf{id}^a, \mathbf{id}^b). \tag{5}$$

In the above Eq. 5, $\mathbf{d}_{q_t} \in \mathbb{R}^D$ denotes the channel embedding of its one-hot identity encoding $\mathbf{id}$, which indicates the speaker role (a or b) of token q_t. The positional encoding is represented as $\mathbf{p}_{q_t} \in \mathbb{R}^D$ indicating which time step both tokens are from. It is important to note that both q_t^a and q_t^b share the same positional embedding, with the Llama 3 (Dubey et al., 2024) model utilizing the Rotary positional embedding as described in (Su et al., 2024b). After the token-pair embedding layer, we obtain the input embedding $\mathbf{e}_{q_t} = [\mathbf{z}_{q_t}, \mathbf{p}_{q_t}, \mathbf{d}_{q_t}]$ for each token q_t (a or b). Following the

implementation of Llama 3, we add both positional embedding and channel embedding to the query and key vectors (instead of value vectors) of each token pair as follows:

$$\mathbf{q} = \mathbf{W}_Q[\mathbf{z}_{q_t^a}, \mathbf{z}_{q_t^b}] + [\mathbf{p}_{q_t^a}, \mathbf{p}_{q_t^b}] + [\mathbf{d}_{q_t^a}, \mathbf{d}_{q_t^b}],\ \mathbf{k} = \mathbf{W}_K[\mathbf{z}_{q_t^a}, \mathbf{z}_{q_t^b}] + [\mathbf{p}_{q_t^a}, \mathbf{p}_{q_t^b}] + [\mathbf{d}_{q_t^a}, \mathbf{d}_{q_t^b}]. \quad (6)$$

Following the above Eq. 6, we obtain the query and key matrices for all token pairs, represented as $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{2T \times D}$, which are projected by weight matrices $\mathbf{W}_Q, \mathbf{W}_K$ respectively. Then we separately multiply codebook embedding vectors by $\mathbf{W}_V$ to obtain the value matrices $\mathbf{V} \in \mathbb{R}^{2T \times D}$. Based on these vectors, we conduct the attention computation as follows:

$$\mathbf{O} = \text{SoftMax}((\mathbf{Q}\mathbf{K}^T/\sqrt{D}) \cdot \mathbf{M})\mathbf{V},\ \mathbf{M} \in \mathbb{R}^{2T \times 2T}. \quad (7)$$

The pair-wise causal masking matrix $\mathbf{M} \in \mathbb{R}^{2T \times 2T}$ is used to mask out the entries in the self-attention matrix, preventing each token q_t from attending to future tokens $(q_{t'}, t' > t)$ and simultaneously attending to tokens from another channel at the same time (i.e. q_t^a and q_t^b cannot attend to each other). The final layer output embedding, denoted as $\mathbf{O}^l \in \mathbf{R}^{2T \times 2T}$, is utilized to generate the next-token-pair prediction $(\hat{q}_{t+1}^a, \hat{q}_{t+1}^b)$ for each (q_t^a, q_t^b) via classifications over codebook embedding indices. The total training loss is equal to the sum of cross-entropy loss over all generated token pair predictions and the ground-truth token pairs. The overall modified embedding layers and self-attention layers are illustrated in Figure 2. Certain advanced architectural components present in Llama 3, such as grouped-queries attention and feedforward layers, have been omitted here, as our modifications do not impact them.

3.3 Streaming Inference

In order to simulate a real-time user-assistant communication scenario, our speech LLM **Parrot** should be proficient in conducting conditional inference with streaming user voice input. In this inference setting, one speaker's voice input is provided as the user, and the model is assigned the task of inferring the other audio channel. This creates a situation that resembles a constrained generation problem. If the inference process strictly follows the training process, then the model should predict $\hat{q}_t^b$ immediately after receiving the speaker's voice input q_t^a at time t. However, due to the VQ-VAE audio tokenization mechanism, it's not feasible to receive just a single audio token from the speaker channel during the streaming inference. This is because the VQ-VAE requires a complete audio signal input within a specific time window. Therefore, unlike the training process, we need to determine when the model should start generating spoken responses upon receiving streaming user input audio tokens. Specifically, we adopt a divide-and-conquer approach to the inference process, breaking it down into chunks, each containing a predetermined number of tokens, denoted as λ. Each time the number of user input tokens reaches λ (a chunk of speaker input is given), our model begins to generate predictions until the number of predicted tokens also reaches λ (a chunk is filled). This procedure is repeated until the end of user voice inputs (e.g., the conclusion of the voice-assistant service). This inference process is illustrated in the accompanying Figure 3.

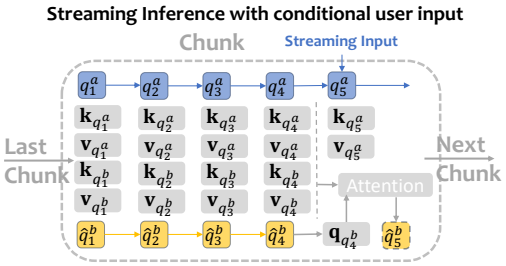


Figure 3: The figure illustrates the chunk-wise streaming inference process. Within each chunk, $(q_1^a, q_2^a, q_3^a, q_4^a, q_5^a)$ represents the provided speaker sequence. Their corresponding keys and values are stored in the KV-cache. **Parrot** sequentially predicts tokens $(\hat{q}_1^b, \hat{q}_2^b, \hat{q}_3^b, \hat{q}_4^b, \hat{q}_5^b)$ based on generated query vectors, which are directed to the Key-Value (KV) cache through attention computations. Once a chunk is filled, the inference process proceeds to the next chunk.

4 Experiments

This section presents the foundational capability evaluation results for **Parrot**. We first describe the two-stage training dataset, data processing methods, and hyper-parameters. We then evaluate the **Parrot**'s performance on core tasks like spoken interaction and provide several case examples.

4.1 DATASET

Parrot employs a two-stage training process. In the first stage, to establish foundational speech capabilities, we trained the model using three speech datasets totaling approximately 14,000 hours. This stage focuses on both speech understanding and synthesis. Unlike other models (Fang et al., 2024) that require audio to be transcribed into text, our **Parrot** only needs single-channel audio for direct training. This reduces the data requirements and, consequently, increases the amount of training data available. For the second stage, we need the **Parrot** to simultaneously gain the ability to listen and speak. To achieve this, we further utilize the Fisher dataset (Cieri et al., 2004). This dataset comprises 2200 hours of phone conversations between randomly paired participants, each discussing a given topic. A notable feature of the Fisher dataset is that each side of the conversation is recorded on separate channels, which allows us to provide ground-truth separated streams to **Parrot**. The original audio is sampled at 8kHz, and we use Librosa [2] to upsample it to 16kHz.

4.2 BASELINES

We compare against baselines from the audio language modeling literature, in three settings. The first category encompasses audio-only models starting from a random initialization, including dGSLM(Nguyen et al., 2023). The second category encompasses several newly released speech LLMs(Zhang et al., 2023a; Xie & Wu, 2024; Fang et al., 2024). As a way to measure the impact of two stage training on spoken fluency, we compare these baselines with **Parrot** trained with and without pre-training phase.

Table 1: The datasets and their usage for training **Parrot**.

Type	Stages	Dataset	Hours
English Reading speech	1	LibriSpeech (Panayotov et al., 2015)	1,000 h
Pronunciation recording	1	Common Voice (Ardila et al., 2019)	3,554 h
Video audio	1	Gigaspeech (Chen et al., 2021)	10,000 h
Spoken English audio	1	Libri-light (Kahn et al., 2020)	60,000 h
Recorded telephone conversation	2	Fisher dataset (Cieri et al., 2004)	2,000 h
Speech Instruction	2	InstructS2S-200K(Fang et al., 2024)	100 h

4.3 TRAINING DETAILS

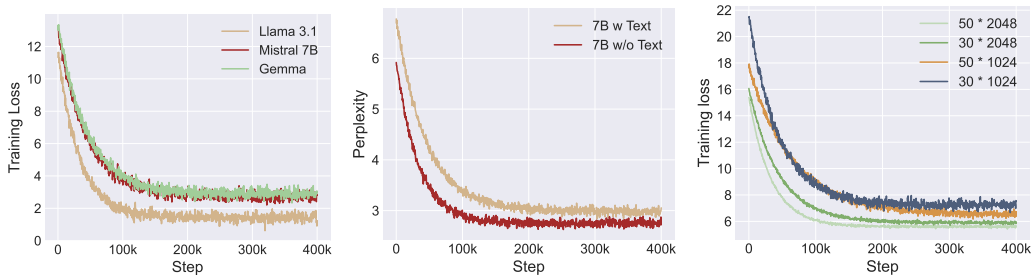
Large Language Model: In this study, we conceptualize audio as an additional language and employ three of the most widely recognized open-source LLMs as our foundational models: Llama-3.1-8B(Dubey et al., 2024), Mistral-7B-v0.3(Jiang et al., 2023a), and Gemma-2-9B(Team et al., 2024). Each of these models comprises an embedding layer, multiple transformer blocks, and a language model (LM) head layer. They all encode the relative positional information of tokens using rotary positional encoding (Su et al., 2024a). **Audio Tokenizer**: We train an audio tokenizer based on (van den Oord et al., 2017), which encodes each second of audio into 30-50 discrete tokens from a codebook of size 2048.

4.4 PRETRAIN EVALUATION

Single audio channel language modeling: We begin by evaluating the capability of **Parrot** to model speech sequences through next-token prediction on the large-scale single channel audio dataset. We use perplexity on the test set's single-channel audio as the metric. The 4a presents the training loss over steps for three distinct models. All three models exhibit a decreasing trend in training loss, indicating effective learning over time. Mistral 7B and Gemma demonstrate similar training loss curves. Notably, Llama 3.1, which exhibits superior text reasoning capabilities, achieves a lower training loss more rapidly compared to Mistral 7B and Gemma. This observation supports our hypothesis that stronger text models can be more effectively adapted to audio tasks, aligning with the conceptualization of "**audio as a new language**".

[2] https://librosa.org/doc

We also explore the trade-off between token rate and codebook size to optimize streaming interaction performance in Figure 4c. Notably, the configuration of 30 * 2048, which represents our chosen compromise solution, demonstrates a balanced performance with a steady decline in training loss.



(a) Training loss curve of different foundation models.

(b) The effectiveness of pre-training without text.

(c) The impact of token rate and codebook size.

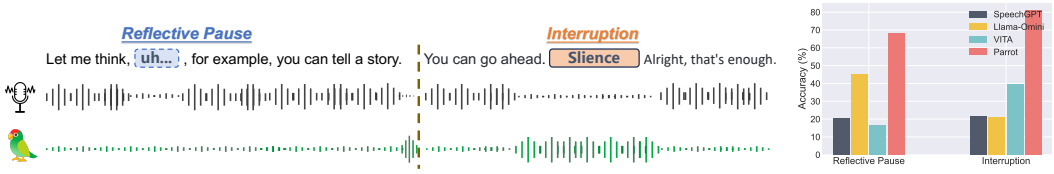
Figure 4: Training loss and perplexity curves for **Parrot** under various Pretraining settings.

4.5 Interactive Evaluation

4.5.1 Reflective Pause and Interruption evaluation

In this section, we use GPT-4 to generate 1,000 text prompts that correspond to two interaction scenarios in natural dialogue: reflective pauses and interruptions. As shown in Figure 5a, reflective pauses evaluate the model's ability to maintain silence during a speaker's contemplative state, while interruptions test the model's response to being interrupted mid-speech. Then we utilize ChatTTS[3] to generate audio prompt corresponding to one of the channels.

Figure 5b compares the accuracy of interaction response by **Parrot** with different baselines. For reflective pauses, **Parrot** demonstrate the highest accuracy at 68%, significantly outperforming the other models. Llama-Omni achieve accuracies of 44%, while SpeechGPT and VITA have an accuracy of only 20% and 19% respectively. Additionally, **Parrot** excelle with an impressive accuracy of 82% for interruption audio prompts, indicating that our model can effectively distinguish human commands.



(a) Interactive evaluation settings. Reflective pauses evaluate the model's ability to maintain silence during a speaker's contemplative state, while interruptions test the model's response to being interrupted by the speaker mid-speech.

(b) Interaction turn-taking event response accuracy.

Figure 5: Reflective pause and interruption evaluation.

4.5.2 Quality and statistics of generated dialogues

We evaluate the linguistic quality and turn-taking dynamics of generated dialogues using various models, as detailed in Table 2. The detailed evaluation settings are in the A.6.2. LSLM(Ma et al., 2024) integrates speaker channels at the embedding layer and separates them in the final layer, demonstrates a notable reduction in the number of Inter-Pausal Units (IPUs) and gaps, indicating smoother transitions between speakers. The dGSLM(Nguyen et al., 2023), particularly with the cross-attention(CA) module, shows a significant decrease in the cumulative duration of pauses and gaps, suggesting more fluid and continuous dialogue. Comparatively, **Parrot** exhibit balanced

[3]https://github.com/2noise/ChatTTS

Table 2: Linguistic quality and turn-taking statistics of generated dialogues, including the number of turn-Taking events and cumulative durations per minute, compared to the ground truth.

Model	Number of occurrences / min				Cumulated duration /min			
	ΔIPU	ΔPause	ΔGap	ΔOverlap	ΔIPU	ΔPause	ΔGap	ΔOverlap
dGSLM w/o CA	-3.9	0.9	-3.6	-1.	-12.1s	8.3s	-1.4s	2.5s
dGSLM	-1.6	3.4	-2.	-2.9	-4.6s	3.6s	0.8s	-1.9s
LSLM	-2.2	3.6	-2.4	-3.2	-4.1s	3.4s	-1.5s	-2.3s
Cascaded	-4.1	-7.	7.4	-6.5	1.3s	-5.5s	0.9s	-3.6s
Parrot$_{0.1}$	-1.4	2.1	-2.0	-1.	-3.2s	**2.5s**	-1.2s	-2.1s
Parrot$_{0.5}$	-1.5	**1.9**	-1.8	-1.5	**-2.9s**	3.0s	**-0.9s**	-2.2s
Parrot$_{0.9}$	**-1.3**	2.2	**-1.5**	**-0.9**	-3.3s	2.8s	-1.4s	**-1.9s**

performance with moderate reductions in both the number and duration of turn-taking events, highlighting their potential for generating natural and coherent dialogues. These findings underscore the importance of model architecture in optimizing dialogue flow and linguistic quality.

4.6 Human Evaluation

We folow the evaluation settings of Veluri et al. (2024) and conduct the evaluation study with 25 annotators with native-level English proficiency. We adapt the Mean Opinion Score (MOS) protocol, utilizing a 5-point Likert scale, to assess the Naturalness (N-MOS) of turn-taking and the Meaningfulness (M-MOS) of dialogue content. Table. 3 compares the meaningfulness and naturalness by **Parrot** with different baselines. Both dGSLM and SyncLLM use Fisher as the only real-world spoken dialogue dataset for training. Besides, we add performance comparison on the out-of-distribution Candor testset(Reece et al., 2023).

Table 3: Meaningfulness (Meaning.) and Naturalness (Nat.) (scores 1-5) mean estimates and standard errors (in parentheses), aggregated overall and for Fisher and CANDOR subsets.

Model	Overall		Fisher		CANDOR	
	Meaning. ↑	Nat. ↑	Meaning. ↑	Nat. ↑	Meaning. ↑	Nat. ↑
dGSLM	1.38 (0.10)	3.85 (0.12)	1.82 (0.09)	4.10 (0.13)	1.51 (0.12)	2.85 (0.18)
SyncLLM	3.85 (0.06)	4.10 (0.05)	4.10 (0.08)	4.33 (0.08)	3.85 (0.09)	3.91 (0.08)
Moshi	3.90 (0.07)	3.95 (0.06)	3.20 (0.10)	4.32 (0.08)	3.90 (0.08)	3.95 (0.08)
Parrot	3.95 (0.04)	4.15 (0.06)	4.10 (0.06)	4.42 (0.06)	4.05 (0.08)	4.05 (0.10)
GT	4.90 (0.01)	4.95 (0.02)	4.90 (0.03)	4.90 (0.04)	4.90 (0.02)	4.95 (0.02)

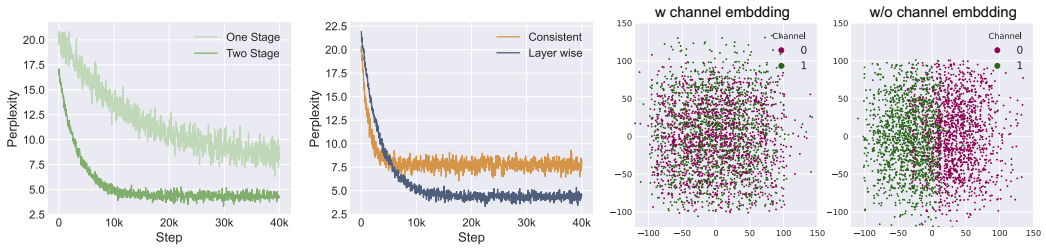
4.7 Ablation Study

In this section, we present an ablation study to evaluate the impact of different channel embedding designs and training stages on the performance of **Parrot**. The results are illustrated in Figure 6.

One-stage VS Two-stage: The Figure 6a compares the perplexity over training steps for these two strategies. The two-stage training approach demonstrates a significantly lower perplexity throughout the training process compared to the one-stage training approach. This indicates that pretraining on single-channel audio data provides a robust foundation, which enhances the model's performance during subsequent fine-tuning on dual-channel data.

Channel Embedding: The Figure 6b illustrates the perplexity over training steps for both approaches. The findings indicate that the layer-wise channel embedding consistently achieves lower perplexity compared to the consistent channel embedding. This suggests that enabling each layer to have its own channel embedding allows the model to learn more effective representations, thereby enhancing performance.

In Figure 6c, we present a t-SNE visualization of token embeddings for each channel, derived from the Fisher testing set conversation. Due to the design of the channel embeddings, there is some separation between the token embeddings from different channels in the reduced-dimensional space



(a) Comparison of perplexity in one-stage and two-stage training strategies.

(b) The impact of channel embedding design.

(c) t-SNE visualization of different channel token embeddings illustrating the distribution of Human (Purple), and Model (Green).

Figure 6: Ablation study on channel embedding designs and training stages.

created by t-SNE. Although there is some overlap between the two channels, these initial findings warrant further exploration and analysis of the embeddings.

5 Case Study

```
Scenario: A user engages in a conversation with Parrot, describing an
    object and asking the model to identify it.
User: Please listen to my description of an object below, and say its
    name when you have guessed it. The description is: it has four legs,
    a flat surface, and is often used for dining or working...
Parrot: I guess it might be a table.
```

Figure 7: Case study of **Parrot** interrupt human speaking correctly and timely.

To intuitively understand the differences in responses from our models, we provide an example in Figure 7. In this scenario, **Parrot** interrupts the user at the precise moment it has gathered enough information to make an accurate prediction. This capability is a significant departure from current models that would typically wait for the user to finish speaking before responding. The ability to interject appropriately not only demonstrates the model's advanced comprehension skills but also enhances the fluidity and naturalness of the interaction.

6 Limitations and Future Works

A current limitation of **Parrot** is its incapacity to integrate the prevalent audio tokenization method, residual vector quantization (RVQ) (Lee et al., 2022). RVQ is typically used to convert continuous audio into discrete tokens, ensuring the preservation of high-quality information. This process involves approximating the audio input with multi-scale tokens, each representing the residual information remaining after the deduction of the previous scale token's information. As a result, the audio token sequence produced by RVQ has an additional residual token dimension (beyond the time step dimension) compared to the standard VQVAE (van den Oord et al., 2017) utilized in **Parrot**. This introduces complexities to the autoregressive generative modeling of spoken dialogue sequences.

7 Conclusion

In conclusion, we introduce a novel spoken dialogue language model, **Parrot**, realized through an innovative pretraining and SFT pipeline. We employ single-channel audio data for pretraining and double-channel audio dialogue data for SFT. To facilitate language modeling on double-channel audio sequences, we unveil the pioneering next-token-pair prediction paradigm for the first time. Comprehensive experiments underscore the superiority of our approach over existing baseline methods. Furthermore, through meticulous ablation studies, we validate the effectiveness of each critical component in our model.

8 ETHICS AND REPRODUCIBILITY STATEMENT

In this study, we propose an innovative spoken dialogue language model, **Parrot**. However, it is important to note that we have not yet conducted a comprehensive safety evaluation of this model. While preliminary results are promising, the potential for unintended consequences, such as biases in audio reasoning or misuse of the technology, remains unassessed. We strongly advocate for further rigorous safety and ethical evaluations to be undertaken by the research community to ensure responsible deployment and to mitigate any adverse impacts.

To ensure the reproducibility of our results, we have made our codebase publicly available through an anonymous git repository, which is provided in the footnote of the abstract. This repository contains comprehensive documentation on data processing, model training, and evaluation procedures, as well as the demo display to facilitate understanding and verification of our methods.

REFERENCES

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse H. Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and Christian Havnø Frank. Musiclm: Generating music from text. *CoRR*, abs/2301.11325, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. Seed-tts: A family of high-quality versatile speech generation models. *CoRR*, abs/2406.02430, 2024.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. 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*, pp. 5723–5738. Association for Computational Linguistics, 2022.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

Siddhant Arora, Hayato Futami, Jee-weon Jung, Yifan Peng, Roshan S. Sharma, Yosuke Kashiwagi, Emiru Tsunoo, and Shinji Watanabe. Universlu: Universal spoken language understanding for diverse classification and sequence generation tasks with a single network. *CoRR*, abs/2310.02973, 2023.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: A language modeling approach to audio generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2523–2533, 2023.

Edresson Casanova, Julian Weber, Christopher Dane Shulby, Arnaldo Cândido Júnior, Eren Gölge, and Moacir A. Ponti. Yourtts: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *International Conference on Machine Learning, ICML 2022, 17-23 July*

2022, Baltimore, Maryland, USA, volume 162 of *Proceedings of Machine Learning Research*, pp. 2709–2720. PMLR, 2022.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 11305–11315. IEEE, 2022.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*, 2021.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 16(6):1505–1518, 2022.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5178–5193. PMLR, 2023.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *CoRR*, abs/2406.05370, 2024.

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. Qwen2-audio technical report. *CoRR*, abs/2407.10759, 2024.

Christopher Cieri, David Graff, Owen Kimball, Dave Miller, and Kevin Walker. Fisher english training speech part 1 transcripts. *Philadelphia: Linguistic Data Consortium*, 2004.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, Zhaocheng Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, Xilai Li, Karel Mundnich, Monica Sunkara, Sundararajan Srinivasan, Kyu J. Han, and Katrin Kirchhoff. Speechverse: A large-scale generalizable audio language model. *CoRR*, abs/2405.08295, 2024.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Trans. Mach. Learn. Res.*, 2023, 2023.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. Technical report, Kyutai, September 2024. URL `http://kyutai.org/Moshi.pdf`.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino,

Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pp. 1–5. IEEE, 2023.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 12873–12883. Computer Vision Foundation / IEEE, 2021.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*, 2024.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Ke Li, Junteng Jia, Yuan Shangguan, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. Audiochatllama: Towards general-purpose speech abilities for llms. 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*, pp. 5522–5532. Association for Computational Linguistics, 2024.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, Ran He, Rongrong Ji, Yunsheng Wu, Caifeng Shan, and Xing Sun. Vita: Towards open-source interactive omni multimodal llm, 2024.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, and Shiliang Zhang. Funasr: A fundamental end-to-end speech recognition toolkit. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pp. 1593–1597. ISCA, 2023.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Défossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. Textually pretrained speech language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3451–3460, 2021.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13916–13932. PMLR, 2023.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, Yi Ren, Yuexian Zou, Zhou Zhao, and Shinji Watanabe. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 23802–23804. AAAI Press, 2024.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023a.

Ziyue Jiang, Yi Ren, Zhenhui Ye, Jinglin Liu, Chen Zhang, Qian Yang, Shengpeng Ji, Rongjie Huang, Chunfeng Wang, Xiang Yin, Zejun Ma, and Zhou Zhao. Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias. *CoRR*, abs/2306.03509, 2023b.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7669–7673, 2020. `https://github.com/facebookresearch/libri-light`.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Rivière, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. Text-free prosody-aware generative spoken language modeling. 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*, pp. 8666–8681. Association for Computational Linguistics, 2022.

Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Trans. Assoc. Comput. Linguistics*, 11:1703–1718, 2023.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pp. 1106–1114, 2012.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021. doi: 10.1162/tacl_a_00430. URL `https://aclanthology.org/2021.tacl-1.79`.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 11513–11522. IEEE, 2022.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *CoRR*, abs/2406.11838, 2024.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*, 2022. doi: 10.1101/2022.07.20.500902.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474. PMLR, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023b.

Ziyang Ma, Yakun Song, Chenpeng Du, Jian Cong, Zhuo Chen, Yuping Wang, Yuxuan Wang, and Xie Chen. Language model can listen while speaking. *CoRR*, abs/2408.02622, 2024.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, James S. Fraser, and Nikhil Vijay Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, pp. 1–8, 2023. URL https://api.semanticscholar.org/CorpusID:256304602.

Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-Weon Jung, Xuankai Chang, and Shinji Watanabe. Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pp. 13326–13330. IEEE, 2024.

Eliya Nachmani, Alon Levkovitch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, R. J. Skerry-Ryan, and Michelle Tadmor Ramanovich. Lms with a voice: Spoken language modeling beyond speech tokens. *CoRR*, abs/2305.15255, 2023.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, R. J. Skerry-Ryan, and Michelle Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered LLM. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. Generative spoken dialogue language modeling. *Trans. Assoc. Comput. Linguistics*, 11:250–266, 2023.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussà, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. Spirit-lm: Interleaved spoken and written language model. *CoRR*, abs/2402.05755, 2024.

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

OpenAI. 2024. URL https://openai.com/index/hello-gpt-4o/.

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

Ajay Patel, Bryan Li, Mohammad Sadegh Rasooli, Noah Constant, Colin Raffel, and Chris Callison-Burch. Bidirectional language models are also few-shot learners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. Speech resynthesis from discrete disentangled self-supervised representations. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pp. 3615–3619. ISCA, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

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

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. Speechbrain: A general-purpose speech toolkit. *CoRR*, abs/2106.04624, 2021.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):eadf3197, 2023.

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, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara N. Sainath, Johan Schalkwyk, Matthew Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirovic, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Havnø Frank. Audiopalm: A large language model that can speak and listen. *CoRR*, abs/2306.12925, 2023.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572–1583, 2019.

Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving AI tasks with chatgpt and its friends in hugging face. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. Graphaf: a flow-based autoregressive model for molecular graph generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024a.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024b. doi: 10.1016/J.NEUCOM.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *CoRR*, abs/2305.16355, 2023.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 3104–3112, 2014.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Raphael Tang, Karun Kumar, Gefei Yang, Akshat Pandey, Yajie Mao, Vladislav Belyaev, Madhuri Emmadi, G. Craig Murray, Ferhan Ture, and Jimmy Lin. Speechnet: Weakly supervised, end-to-end speech recognition at industrial scale. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: EMNLP 2022 - Industry Track, Abu Dhabi, UAE, December 7 - 11, 2022*, pp. 285–293. Association for Computational Linguistics, 2022.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *CoRR*, abs/2405.09818, 2024.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *CoRR*, abs/2404.02905, 2024.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf.

Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. Beyond turn-based interfaces: Synchronous llms as full-duplex dialogue agents, 2024. URL https://arxiv.org/abs/2409.15594.

Jiaming Wang, Zhihao Du, Qian Chen, Yunfei Chu, Zhifu Gao, Zerui Li, Kai Hu, Xiaohuan Zhou, Jin Xu, Ziyang Ma, Wen Wang, Siqi Zheng, Chang Zhou, Zhijie Yan, and Shiliang Zhang. Lauragpt: Listen, attend, understand, and regenerate audio with GPT. *CoRR*, abs/2310.04673, 2023a.

Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. Viola: Unified codec language models for speech recognition, synthesis, and translation. *CoRR*, abs/2305.16107, 2023b.

Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Haibin Wu, Kai-Wei Chang, Yuan-Kuei Wu, and Hung-yi Lee. Speechgen: Unlocking the generative power of speech language models with prompts. *CoRR*, abs/2306.02207, 2023. doi: 10.48550/ARXIV.2306.02207. URL https://doi.org/10.48550/arXiv.2306.02207.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal LLM. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation, 2024.

Zhifei Xie and Changqiao Wu. Mini-omni: Language models can hear, talk while thinking in streaming, 2024. URL `https://arxiv.org/abs/2408.16725`.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:1720–1733, 2023.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. Instructtts: Modelling expressive TTS in discrete latent space with natural language style prompt. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:2913–2925, 2024.

Jiaxuan You, Rex Ying, Xiang Ren, William L. Hamilton, and Jure Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5694–5703. PMLR, 2018.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30: 495–507, 2022.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pp. 15757–15773. Association for Computational Linguistics, 2023a.

Dong Zhang, Xin Zhang, Jun Zhan, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechgpt-gen: Scaling chain-of-information speech generation. *CoRR*, abs/2401.13527, 2024a.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024b.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara N. Sainath, Pedro J. Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. Google USM: scaling automatic speech recognition beyond 100 languages. *CoRR*, abs/2303.01037, 2023b.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model, 2024.

Xiaohuan Zhou, Jiaming Wang, Zeyu Cui, Shiliang Zhang, Zhijie Yan, Jingren Zhou, and Chang Zhou. Mmspeech: Multi-modal multi-task encoder-decoder pre-training for speech recognition. In *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, pp. 4943–4947. ISCA, 2023.

A Appendix

A.1 Detailed Related Work Discussions

We compare **Parrot** with several newly released speech LLMs, which are Mini-Omni (Xie & Wu, 2024), Llama-Omni (Fang et al., 2024), Moshi (Défossez et al., 2024), LSLM (Ma et al., 2024).

1) Mini-Omni: The major advancement of this model is the batched parallel decoding strategy.
   - Advantages: Text generation can significantly enhance the quality of the audio produced. Concurrently, the implementation of batched parallel decoding can substantially mitigate issues related to inference latency. Overall, Mini-Omni effectively maintains a high standard of response quality while circumventing the latency typically associated with TTS translations.
   - Limitations: This model, while a multi-modal QA system, adheres to the standard architecture of multi-modal LLMs with various modality adaptors. However, it falls short in handling natural spoken conversations with real-time streaming user voice inputs. The dynamic nature of real-time dialogues, characterized by various pauses and turn-taking events, cannot be effectively simulated by this system.
2) Llama-Omni: This speech LLM also mainly focuses on enhancing the decoder stage like the previous Mini-Omni model. It propose an non-autoregressive decoder to simultaneously generate texts and audios. The text token is firstly upsampled and then fed into the speech decoder to derive the output voice. Unlike traditional TTS, Llama-Omni applies TTS word by word in an non-autoregressive manner.
   - Advantages: Like the Mini-Omni, this model also enjoys the response reliability due to the usage of intermediate text generation. In this way, Llama-Omni also enjoys low inference latency while maintaining high-quality content response.
   - Limitations: The Llama-Omni also shares the same limitations like Mini-Omin. Relying on text generations cannot handle special speech tokens that are hard to match to text tokens. In addition, the multi-modal LLMs can only handle multi-turn QA while failing to handle natural conversations like interruptions and pauses.
3) LSLM: This speech LLM explicitly leverages the double-channel audio data. Unlike **Parrot**, LSLM fuses two channel tokens into one single token and still follows the next-token prediction training objective. To enable LSLM to learn to interrupt, this work trains the speech LLM on the synthetic interruption data.
   - Advantages: No need to change the next-token prediction paradigm of the original LLM, which keeps the speech LLM as simple as possible.
   - Limitations: The introduction of the special "EOS" token and the "interruption" token will bring additional challenges in audio preprocessing. A threshold must be determined to filter what tokens are assigned to be "interruption token", which can be tricky. In addition, this model can only learn to interrupt by training on specific synthetic data. First, it might be troublesome to synthesize turn-taking events. Second, there is always a distribution gap between synthetic turn-taking and real-world turn-taking.
4) Moshi: This is a newly open-sourced speech LLM with high-quality spoken responses and minimal inference latency. Moshi leverages the RVQ technique to tokenize the audio inputs. And it explicitly proposes the usage of multi-channel audio modeling. There are mainly text channels, speaker audio channels and listener audio channels. The generative modeling of the multi-channel token sequences is following the RQtransformer (Lee et al., 2022), which is an encoder-decoder architecture.
   - Advantages: The usage of RVQ can largely improve the quality of discrete audio representations. And the usage of intermediate text translation can significantly improve the reliability of response contents.
   - Limitations: The multi-channel data structure requires the alignment between text sequences and audio sequences, which is a non-trivial engineering work. Also, the encoder-decoder RQtransformer architecture requires to receive the entire input of speaker's channel, which still somehow downgrades the modeling efficency. Last but not least, this model can be regarded as alternative form of online cascading approach, which relies on the accuracy of both audio-to-text and text-to-audio generation.

In comparison to the above models, **Parrot** enjoys several important advantages:

- **Real Streaming Inference: Parrot** is capable of managing real-time streaming inference, eliminating the need for specific training on turn-taking, as required by models like LSLM. It can interact seamlessly with human users through natural turn-taking for the duration of

the service. In contrast, multi-modal speech LLMs such as Mini-Omni and Llama-Omni can only interact with users on a turn-by-turn basis. In essence, **Parrot** does not depend on manually-defined interruption rules when conducting streaming inference.

- **Decoder-only Transformers:** In contrast to the encoder-decoder dialogue language modeling, **Parrot** employs a decoder-only transformer. This architecture offers numerous significant advantages. For instance, the encoder-decoder structure necessitates maintaining a window to receive complete inputs during the inference stage. However, the decoder-only architecture simply requires querying the cached key-value pairs, resulting in superior computational efficiency during inference.
- **Spoken Dialogue Data Usage Efficiency:** Both Moshi and LSLM randomly assign one channel as the speaker and another as the listener. This approach potentially reduces dialogue data efficiency, as the trained model becomes speaker-dependent. Essentially, the model needs to train the reverse conditional distribution by swapping the roles, which could pose scalability issues as more channels are added in the future. In contrast, **Parrot** is speaker-independent and concurrently learns the conditional distribution of both speaker's audio channels.

A.2 Representation of the joint sequence and mask strategy modeled by **Parrot**

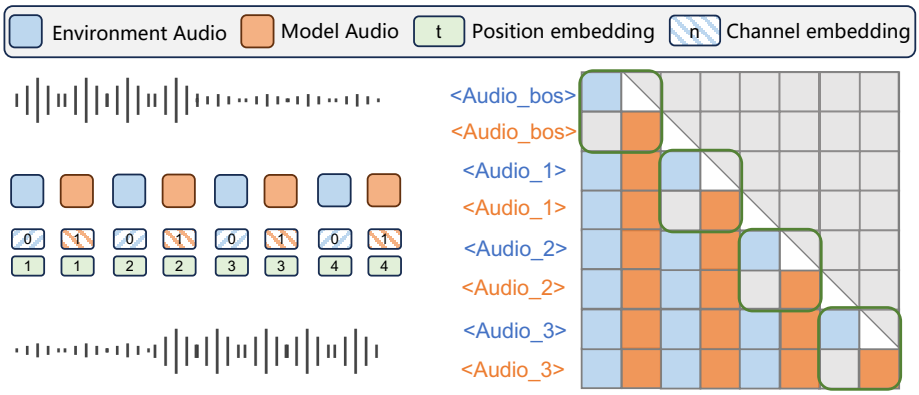


Figure 8: Pair-wise causal masking attention for next-token-pair prediction.

A.3 Potential Solutions to Limitations of **Parrot**

To overcome the limitations previously discussed, we propose a potential solution: the creation of a novel generative model for RVQ-based dual-channel audio sequences. However, the complexity of this task is heightened due to the unclear dependency relations across two distinct dimensions - the time dimension and the residual token dimension. As an alternative, we could opt to refine our method by increasing the number of discrete tokens per second. This approach would circumvent the need for RVQ while simultaneously enhancing the quality of the audio information. In future research, our goal is to train our method on substantially larger datasets and concurrently develop more sophisticated speech language model architectures. We hypothesize that the performance of our method can be further elevated to a new level through various potential approaches, without the direct application of RVQ.

A.4 More Implementation Details and Hyper-parameter Settings

A.4.1 Hyper-parameter Settings

Our model is trained on 16 A100 GPUs, utilizing a cosine annealing learning rate scheduler with a minimum learning rate of 4e-6 and a maximum learning rate of 4e-4. Each training epoch consists of 40,000 steps, with batch size 192 for each step. During fine-tuning, we use learn rate from 4e-6 to 5e-5.

A.4.2 STREAM INFERENCE

Table 4: Latency, speech-text alignment and speech quality under different unit chunk sizes.

Chunk Size Ω	Latency (ms)	#Lagging Word	ASR-WER ↓	ASR-CER ↓
10	310	2.1	12.5	7.42
20	320	3.1	12.65	7.45
40	350	4.4	12.45	7.89
60	410	6.9	13.10	8.10
80	490	10.2	14.50	8.35
100	550	11.3	15.30	9.05

A.5 CASE STUDY

Here, we present several cases to demonstrate **Parrot**'s capabilities in speech understanding and reasoning.

A.6 MORE EXPERIMENTAL RESULTS

A.6.1 AUDIO TOKENIZER QUALITY

Table 5: Comparison of different models and tokenizers on objective and subjective metrics.

Model	Tokenizer	Objective		Subjective	
		WER↓	SIM↑	MOS↑	SMOS↑
Groundtruth		1.9	0.93	4.5	3.96
VALL-E	EnCodec	7.9	0.75	3.08	3.31
USLM	SpeechTokenizer	7.2	0.81	3.63	3.45
Parrot	VQVAE	**6.9**	**0.82**	**3.71**	**4.50**

A.6.2 DIALOGUE LINGUISTIC QUALITY

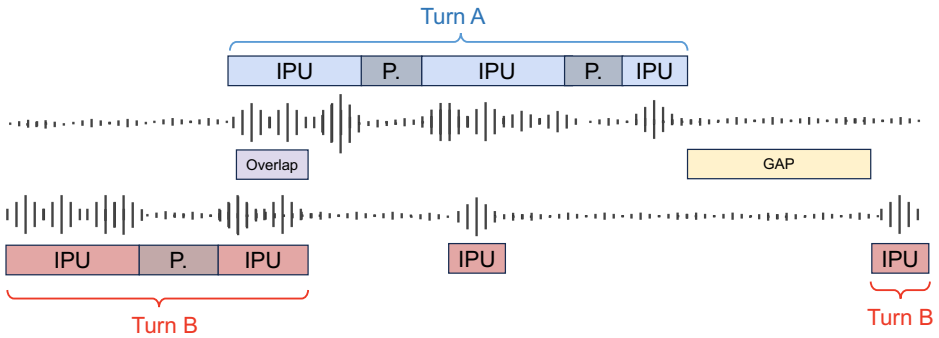


Figure 9: Illustration of turn-taking events: IPU (Interpausal Unit), Turn (for speaker A and Speaker B, resp), P.(within-speaker Pause), Gap and Overlap.

Our model generates two audio channels at the same time, allowing us to use basic Voice Activity Detection (VAD) tools on the output to gather turn-taking metrics. According to the settings in (Nguyen et al., 2023), an Inter-Pausal Unit (IPU) is a continuous speech segment within one speaker's channel, bordered by VAD-detected silences longer than 200ms on both ends. Silence is defined as the lack of voice signals on either channel, while overlap refers to segments where voice signals are detected on both channels. Silences can be further divided into gaps (between IPUs of

different speakers) and pauses (within the same speaker's IPUs). Consecutive IPUs by the same speaker, separated by a pause, are merged into a single turn. Our analysis will focus on measuring the duration distribution of IPUs, gaps, pauses, and overlaps in both the training corpus and the dialogues generated by our various models.

A.6.3 REFLECTIVE PAUSE AUDIO DATASET

Prompt for reflective pause

"Hmm..., this question is a bit complicated, I need to think about it."
"Let me recall, uh..., yes, we went to the park that day."
"You know, that..., oh, yes, it's the new restaurant."
"I remember he mentioned it, um..., it seems to be last Friday."
"This matter, um..., I think we need to discuss it again."
"Let me think about it, uh..., yes, that's it."
"I'm not sure, um..., maybe I need to confirm it again."
"This question, um..., I think we can solve it this way."
"Let me think about it again, uh..., yes, I remember it."
"The one you mentioned, um..., I seem to have some impression."
"We need to deal with the budget issue of this project. Um..., this problem is a bit complicated, I need to think about it."
"Do you remember the last time we met? Let me recall, uh..., yes, we went to the park that day."
"Have you heard about the new restaurant? You know, that..., oh, yes, that new restaurant."
"When did he tell you the news? I remember he mentioned it, uh..., it seems to be last Friday."
"Do you have any suggestions about this plan? This matter, uh..., I think we need to discuss it again."
"Can you give me an example? Let me think about it, uh..., yes, that's it."
"Are you sure this data is correct? I'm not sure, uh..., I may need to confirm it again."
"How should we deal with this emergency? This problem, uh..., I think we can solve it this way."
"Can you explain this concept again? Let me think about it again, uh..., yes, I remember it."
"Do you know what he is talking about? The one you said, uh..., I seem to have some impression."

Prompt for GPT score

Content (1-5 points):
1 point: The response is largely irrelevant, incorrect, or fails to address the user's query. It may be off-topic or provide incorrect information.
2 points: The response is somewhat relevant but lacks accuracy or completeness. It may only partially answer the user's question or include extraneous information.
3 points: The response is relevant and mostly accurate, but it may lack conciseness or include unnecessary details that don't contribute to the main point.
4 points: The response is relevant, accurate, and concise, providing a clear answer to the user's question without unnecessary elaboration.
5 points: The response is exceptionally relevant, accurate, and to the point. It directly addresses the user's query in a highly effective and efficient manner, providing exactly the information needed.

Style (1-5 points):
1 point: The response is poorly suited for speech interaction, possibly including structured elements like lists or being overly complex, disjointed, or difficult to understand.
2 points: The response is somewhat suitable but may be too long, too short, or awkwardly phrased, making it less effective in a speech interaction context.
3 points: The response is generally suitable for speech interaction, but it may have minor issues with length, clarity, or fluency that detract slightly from the overall effectiveness.
4 points: The response is well-suited for speech interaction, with appropriate length, clear language, and a natural flow. It is easy to understand when spoken aloud.
5 points: The response is perfectly suited for speech interaction. It is the ideal length, highly clear, and flows naturally, making it easy to follow and understand when spoken.

Below are the transcription of user's instruction and models' response:
[Instruction]: **{instruction}**
[Response]: **{response}**

After evaluating, please output the scores in JSON format: {"content": content score, "style": style score}. You don't need to provide any explanations.

A.7 MOTIVATIONS OF USING DOUBLE-CHANNEL SPOKEN DIALOGUE DATA

Inspired by GPT-4o (OpenAI, 2024), we aspire to create a powerful voice assistant that can engage with human users in a natural and fluent way. Ideally, the assistant should be able to be interrupted by users. If a user needs to convey something urgently, the assistant should stop speaking and listen attentively. Furthermore, when a user is in thought or taking a pause, the assistant should not prematurely conclude that the user has finished speaking. Instead, it should patiently wait for the user to complete their thoughts. An advanced voice assistant could even interrupt users when it has already grasped their intentions, much like how we often interrupt each other in daily conversations. There are numerous other scenarios that an intelligent voice assistant should be equipped to handle. Given these complex application scenarios, it's challenging to address these issues through simple manual engineering, such as the introduction of special tokens like silence tokens, or hard interruptions when the user is speaking.

The success of foundational models hinges on our trust in the model's capacity to learn autonomously from data, rather than over-interfering with the learning process or over-engineering the neural architectures and algorithms. Consequently, in this paper, we utilize double-channel dialogue data and directly train the speech LLM on this spoken dialogue data. With robust pre-trained speech LLMs, we can reasonably anticipate that the model can learn how humans converse with each other by directly "reading" their dialogues. This approach eliminates the need for setting manual rules to assist the voice assistant in scenario judgement. The assistant may learn how to navigate these scenarios by processing a sufficient amount of spoken dialogue data. Regrettably, the current availability of open-source double-channel spoken-dialogue data is limited. Looking ahead, we hope our work will stimulate the community to gather large-scale double-channel or even multi-channel spoken dialogue data.