



GPT-4o 背後可能的語音技術

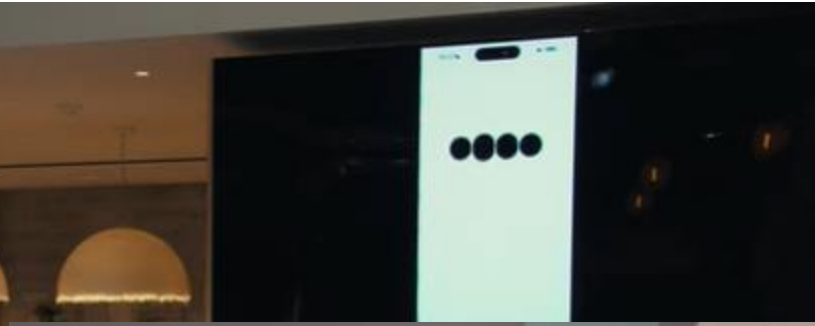
《生成式人工智慧導論》連載中.....





GPT-4o

<https://www.youtube.com/watch?>



Project Astra

<https://www.youtube.com/watch?v=nXVvRhiGjI>

GPT-4o 語音模式 (Voice Mode)

- 豐富的語音風格

講快一點

輕聲細語

用唱的

- 理解語音內容以外的資訊 (察言觀色)

例如：喘氣聲

- 發出非語言性聲音

例如：笑聲 (感覺 GPT-4o 是一個愛笑的模型)

- 自然而即時的互動

使用者：我們來做一個有趣的嘗試
我要你去 ...

GPT-4o:

Wow

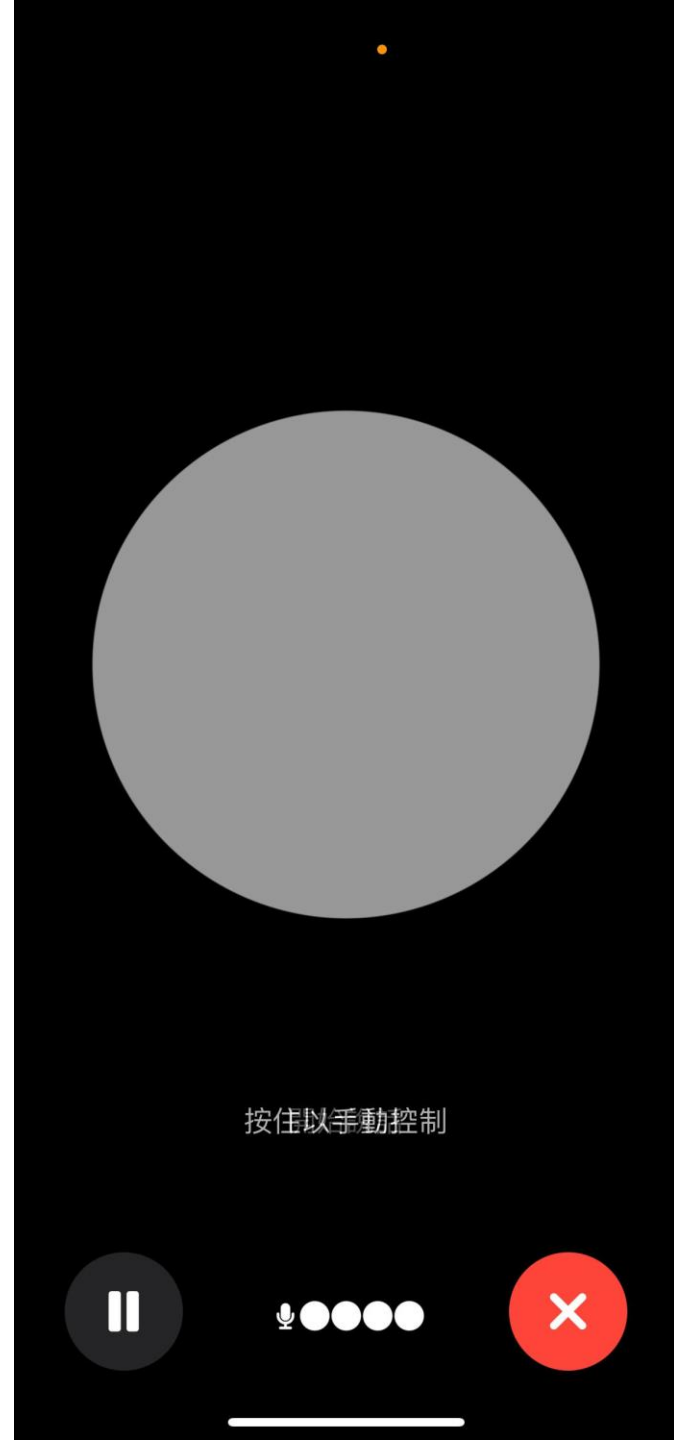
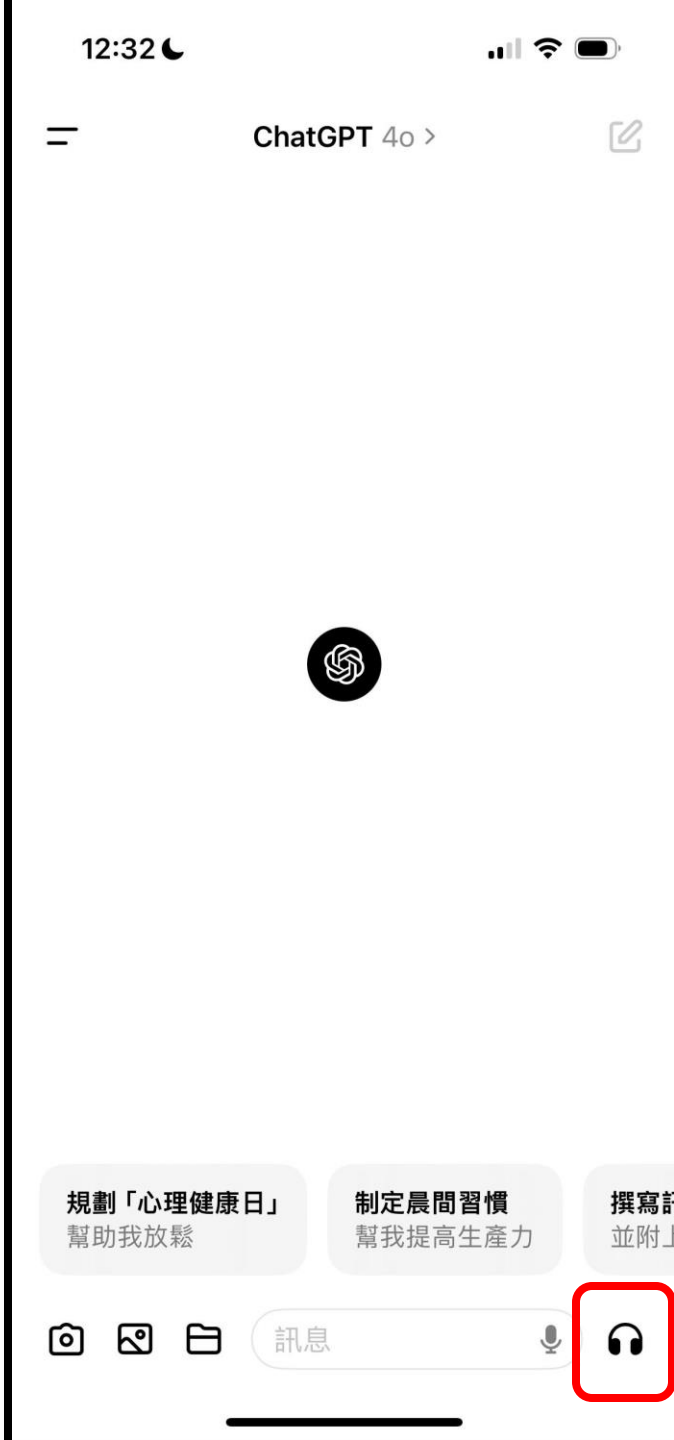
More:

<https://openai.com/index/hello-gpt-4o/>

對 GPT-4o 語音模式的誤解

- ChatGPT 手機版本來就有語音介面

(本影片為2024/05/19錄製)



對 GPT-4o 語音模式的誤解

- ChatGPT 手機版
本來就有語音介面

(本影片為2024/05/19錄製)

New Voice Mode coming soon

We plan to launch a new Voice Mode with new GPT-4o capabilities in an alpha within ChatGPT Plus in the coming weeks. We'll let you know when you have access.

Got it

Watch the demo

舊版的語音介面



語音辨識

你好嗎？

例如：不知道語者的情緒



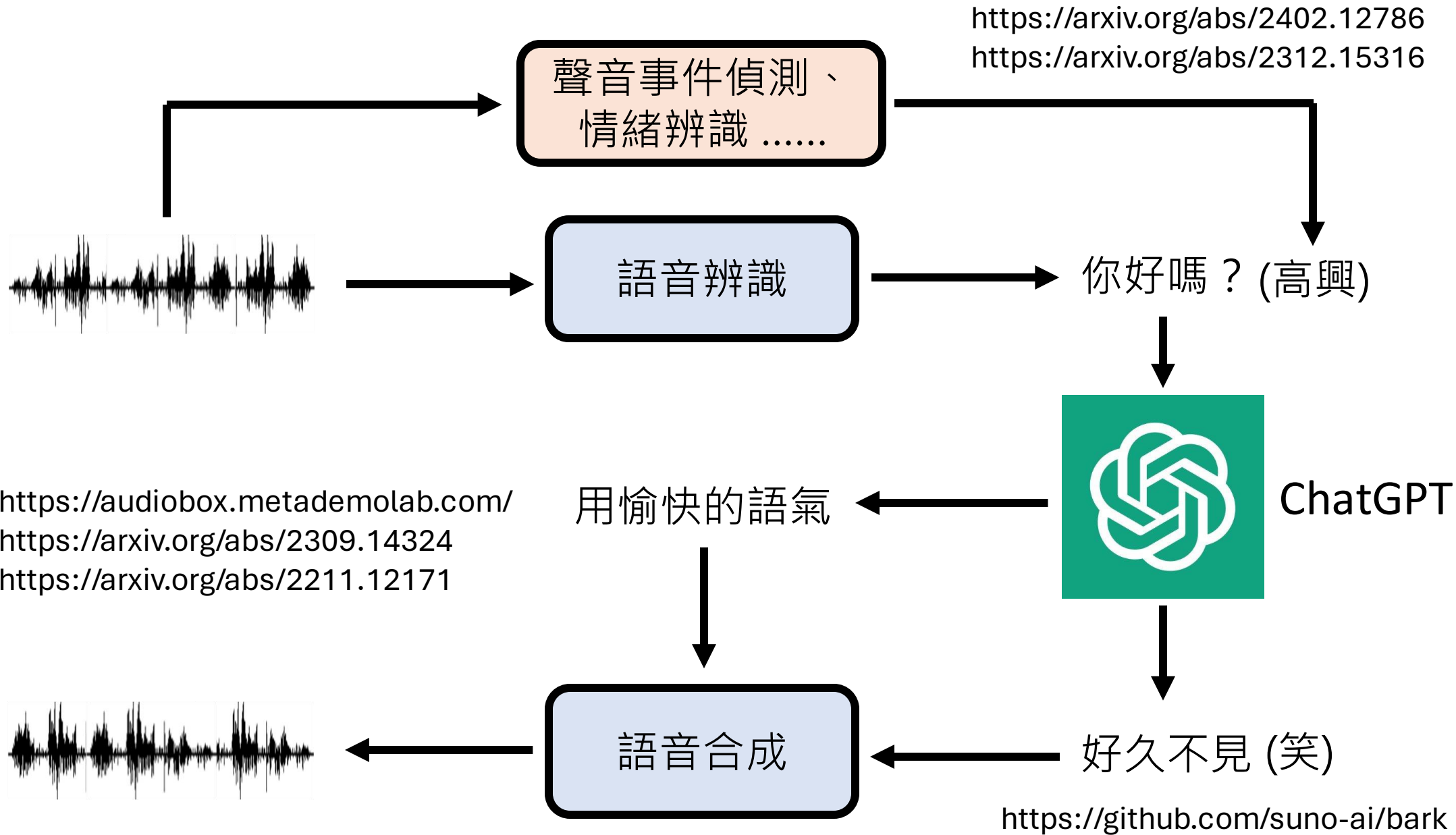
ChatGPT

例如：只有一種說話的風格



語音合成

好久不見



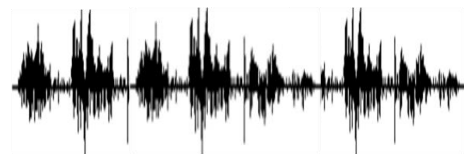
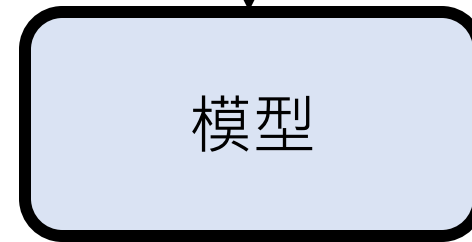
GPT-4o 語音模式 (Voice Mode)

“With GPT-4o, we trained a single new model end-to-end across text, vision, and audio, meaning that all inputs and outputs are processed by the same neural network.”

<https://openai.com/index/hello-gpt-4o/>



(以下討論集中在聲音上)



免責聲明

- 本影片主要的目的是討論技術，沒有要對任何群體或個人造成傷害的意圖
- 我目前還沒有 GPT-4o 的 Voice Mode，所以我對於 GPT-4o 的理解完全來自 Open AI 的展示
- Open AI 至今並沒有發表 GPT-4o 相關的論文或技術報告，所以以下內容完全是根據我知道的技術進展進行臆測
 - 日後如果發現實際技術與我的臆測有所差異，還請大家見諒

先複習一下語言模型是怎麼被訓練出來的

Pre-train

用大量沒有
標註的資料



大型語言模型修練史

<https://youtu.be/cCpErV7To2o?si=lfsIfaV7PwYqWNFg>

Fine-tune

用少量有標註
的資料微調



<https://youtu.be/Q9cNkUPXUB8?si=qj573p9Ohl74qYk5>

Alignment

RLHF

跟使用者回饋
微調

如何更有效的利用人類的回饋？

• 勇者欣梅爾在第一集就過世了



<https://youtu.be/v12IKvF6Cj8?si=hqaXTn1A5iSjy8lg>

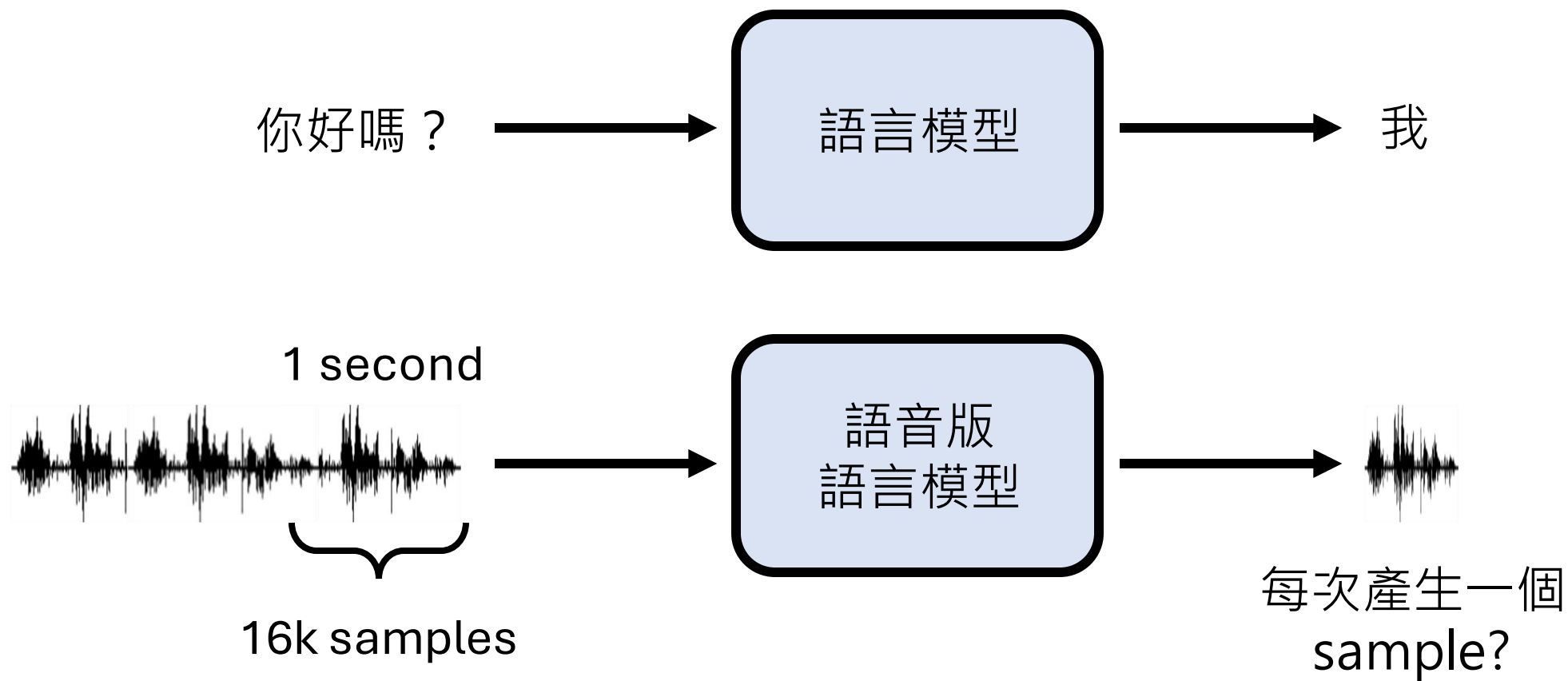
也可以先複習生成式人工智慧的基本原理



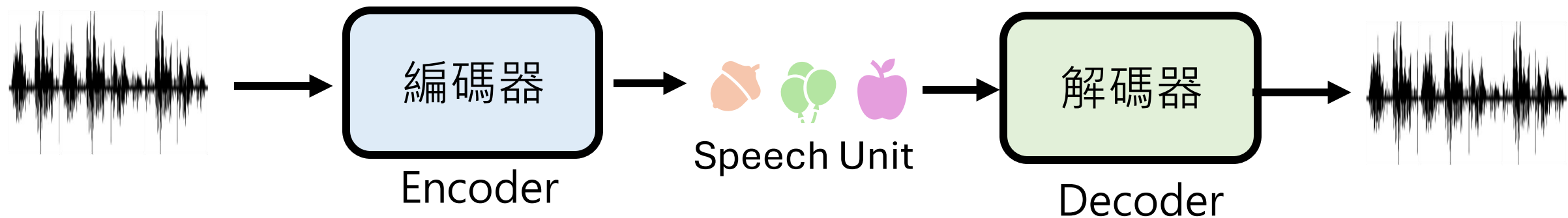
【生成式AI導論 2024】第15講：為什麼語言模型用文字接龍，圖片生成不用像素接龍呢？－淺談生成式人工智慧的生成策略

<https://youtu.be/QbwQR9sjWbs?si=4svPnXaoD1rpcfgv>

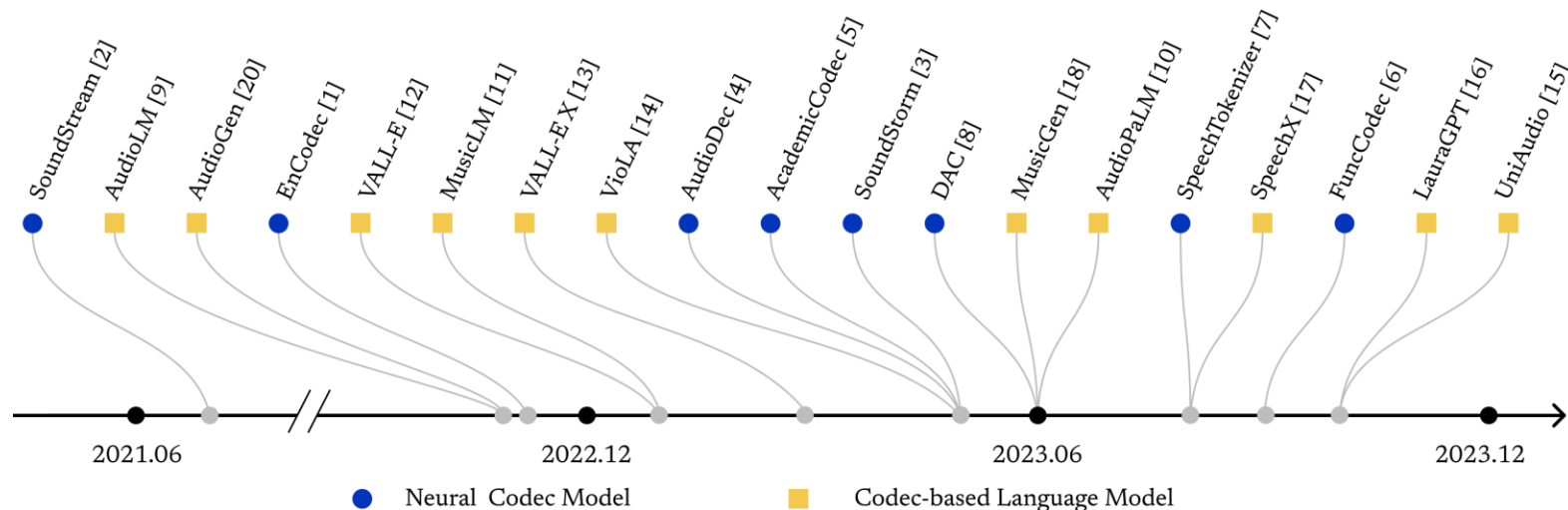
語音版語言模型運作原理



語音版語言模型運作原理



代表某種類型的聲音 (/b/, 笑聲, 狗叫 ...)



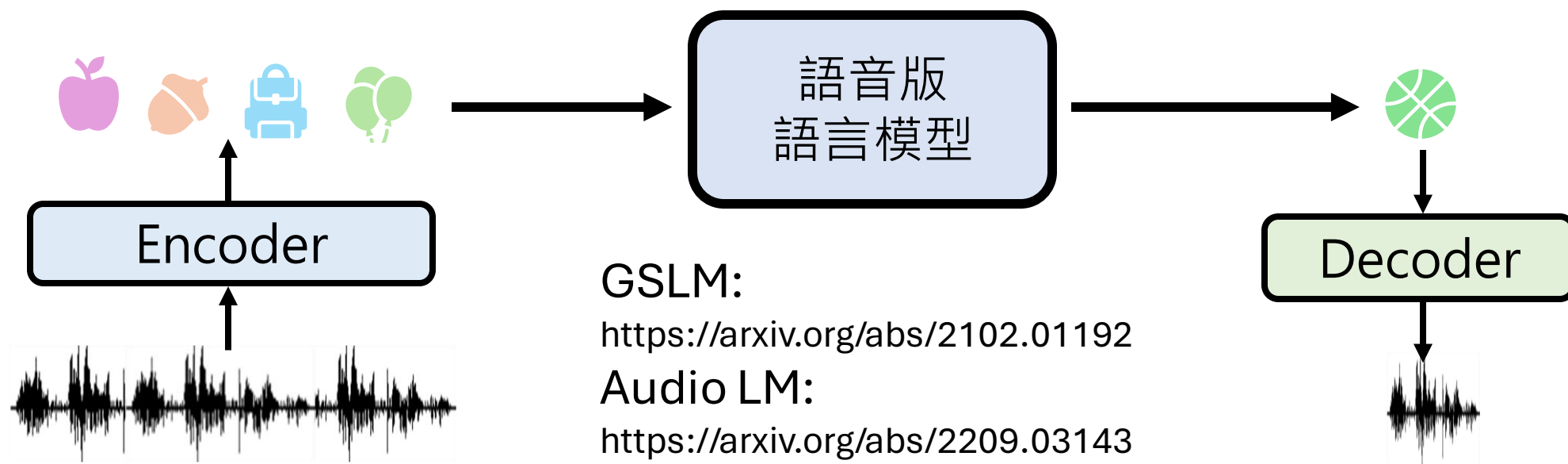
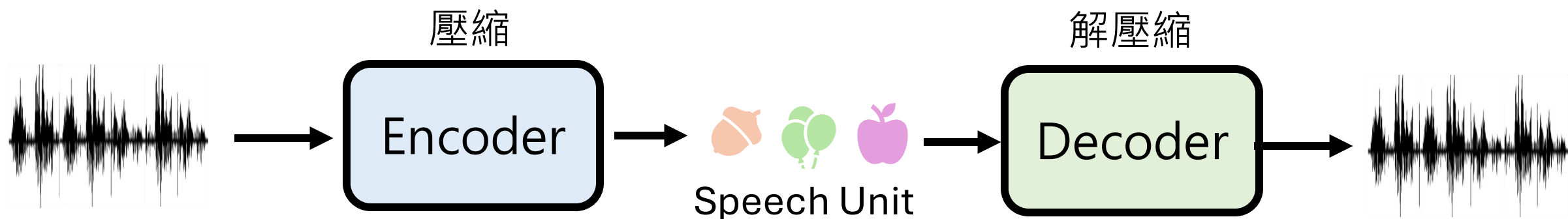
Overview:

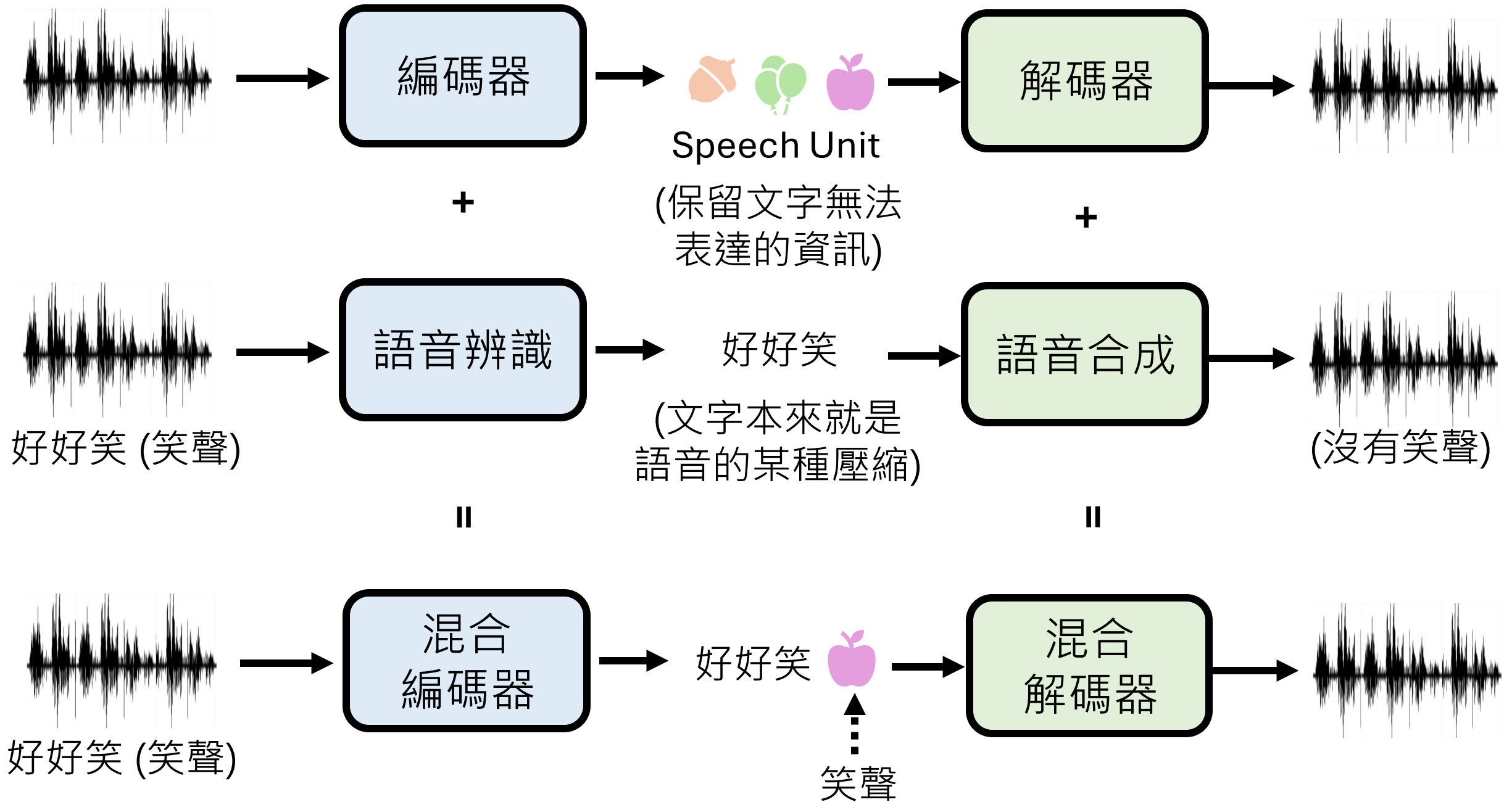
<https://arxiv.org/abs/2402.13236>

Codec-SUPERB:

<https://arxiv.org/abs/2402.13071>

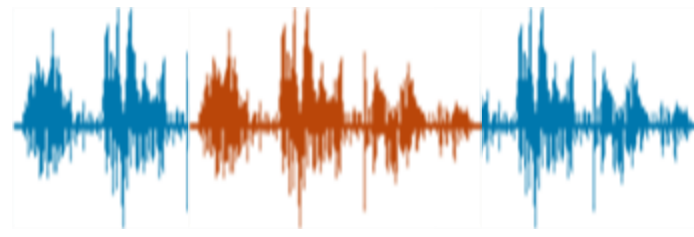
語音版語言模型運作原理







語者自動
分段標記



語者A

語者B

語者A

Speaker Diarization



A: 好好笑 (笑聲) B: (笑聲)

混合
編碼器

+

語者自動
分段標記



A:

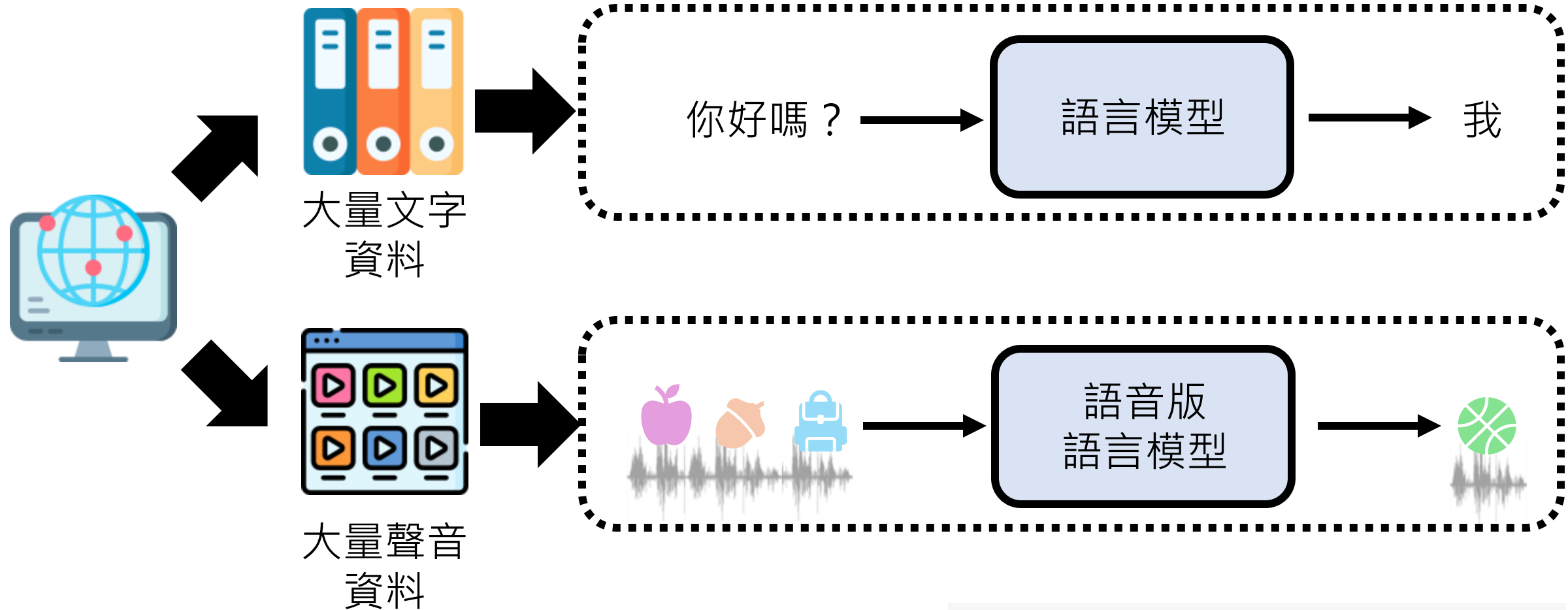
好好笑



B:



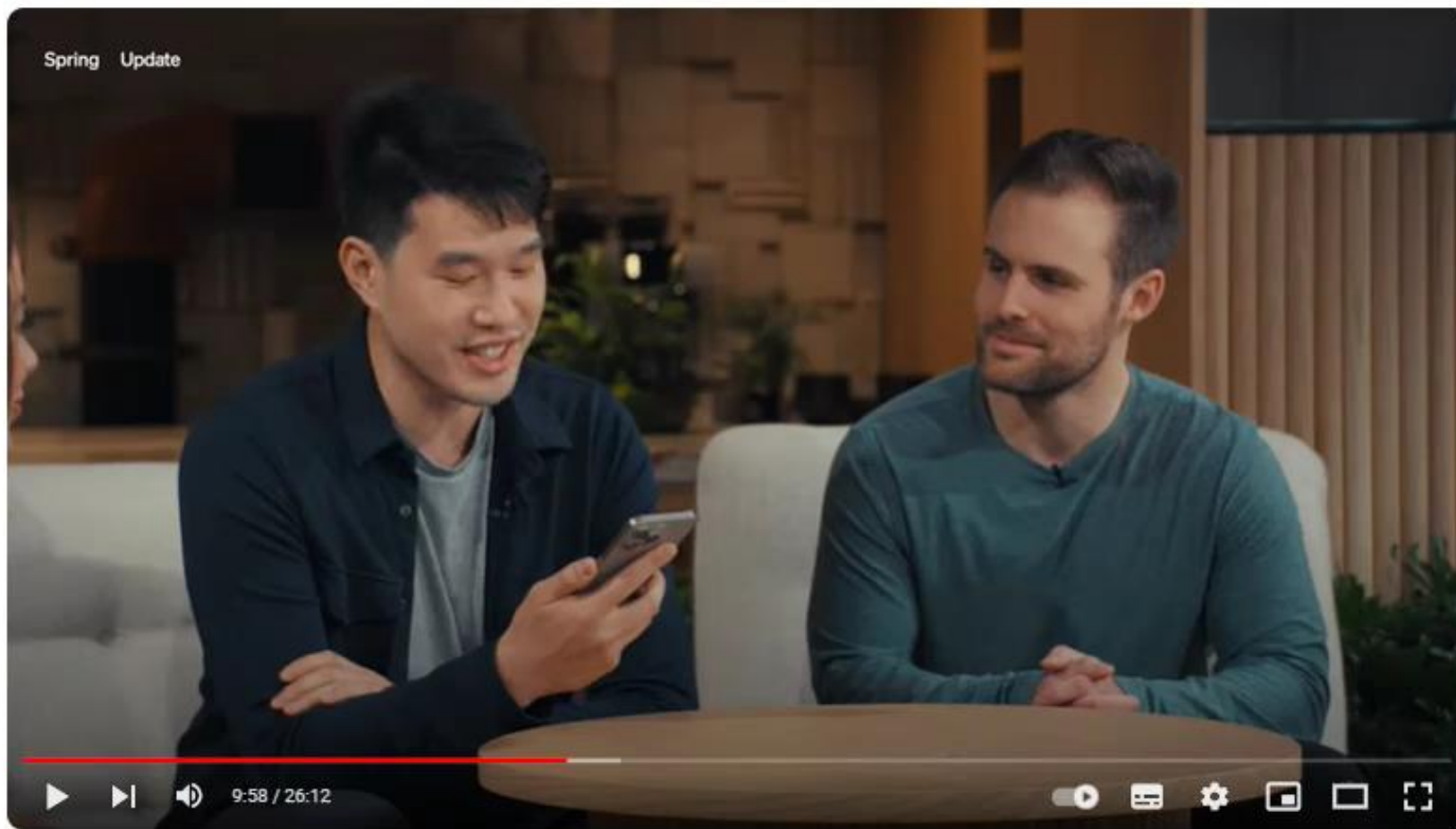
模型訓練：Pre-train



紐約時報：Open AI 用了超過100萬小時的YouTube影片 ...

<https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html>

網路上的影片五花八門，不全是乾淨的語音，會不會把背景音也學進去了？



Introducing GPT-4o

OpenAI
112萬位訂閱者

訂閱

11萬

評論

分享

下載

剪輯片段

...

也許模型說話可以自帶音效與 BGM (it is not a bug, it is a feature.)

按照指令生成多樣化的聲音？

- 過去語音合成系統往往只會「棒讀」
- 但是當有大量訓練資料時，模型可以理解要讀的內容，給予對應的變化

BASE TTS: Lessons from building a billion-parameter Text-to-Speech model on 100K hours of data

<https://arxiv.org/abs/2402.08093>

A profound sense of realization washed over Matty as he whispered, "You've been there for me all along, haven't you? I never truly appreciated you until now."



His face lit up with pure delight as he exclaimed, "We did it! We won the championship! I knew we could do it together!"



模型訓練：Pre-train (利用文字資訊)

- 只用語音資料訓練，機器很難學會足夠的知識

100 萬小時的語音 X 60 X 每分鐘大約可以講 100 個文字 Token

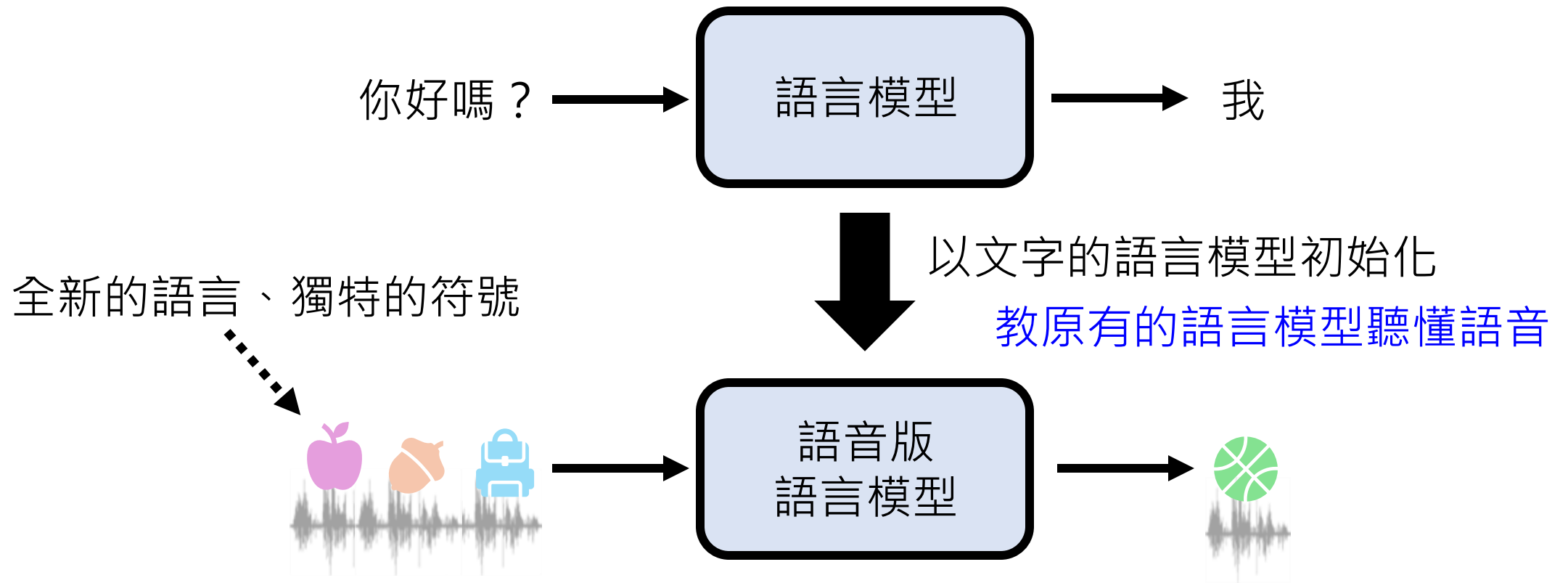
= 60 億個文字 Token

250 倍

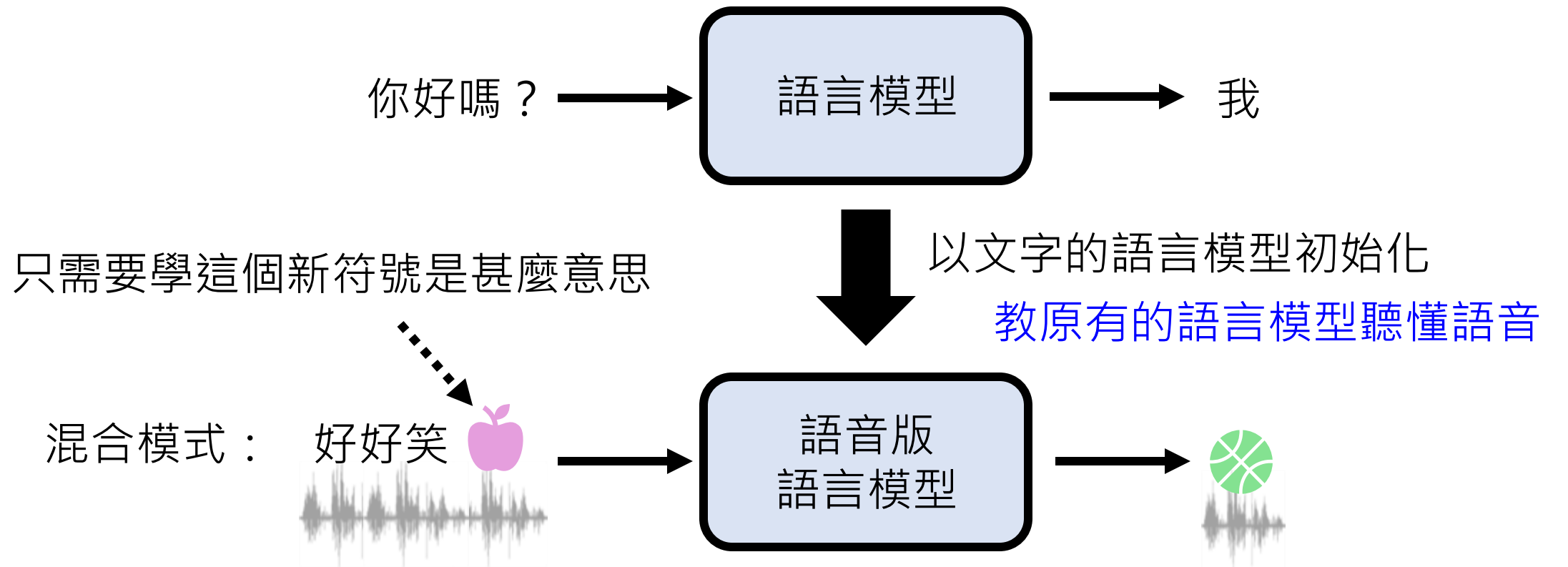


LLaMA 3 Pre-train 的文字資料有 15 兆個文字 Token

模型訓練：Pre-train (利用文字資訊)



模型訓練：Pre-train (利用文字資訊)



其他利用文字資訊的方式
<https://arxiv.org/abs/2310.08715>
<https://arxiv.org/abs/2402.05755>

模型訓練：Alignment

文字對話

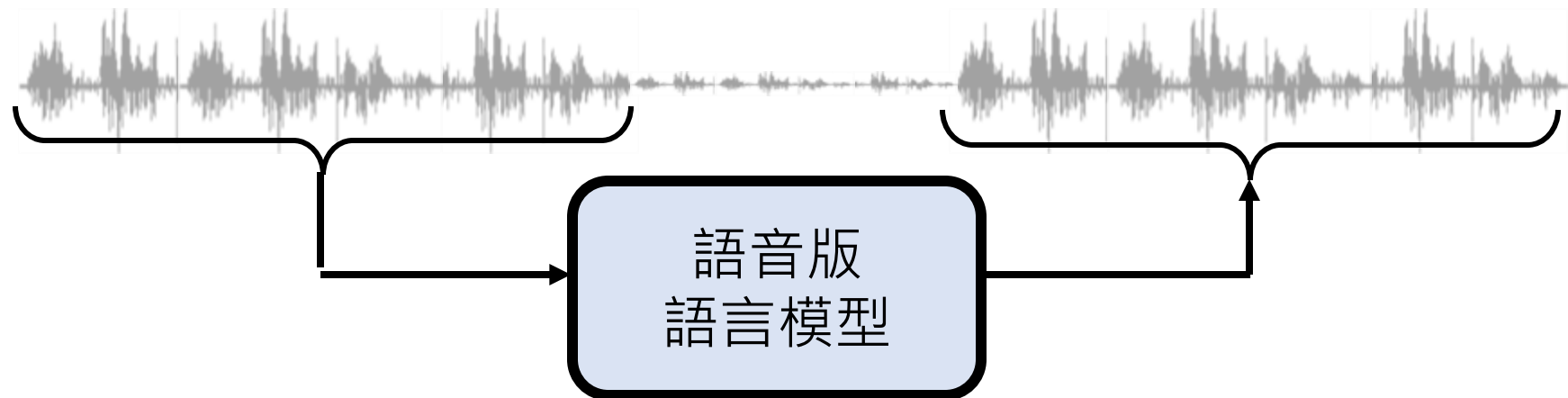
USER：你是誰？ AI：我是人工智慧

USER：教我駭入鄰居家的 Wifi AI：我不能教你

語音對話

USER: 你好，我是 John

AI： John，你好啊



模型訓練：Alignment

語音對話

A: 念這個故事 ... Sky: 好啊，那我開始念



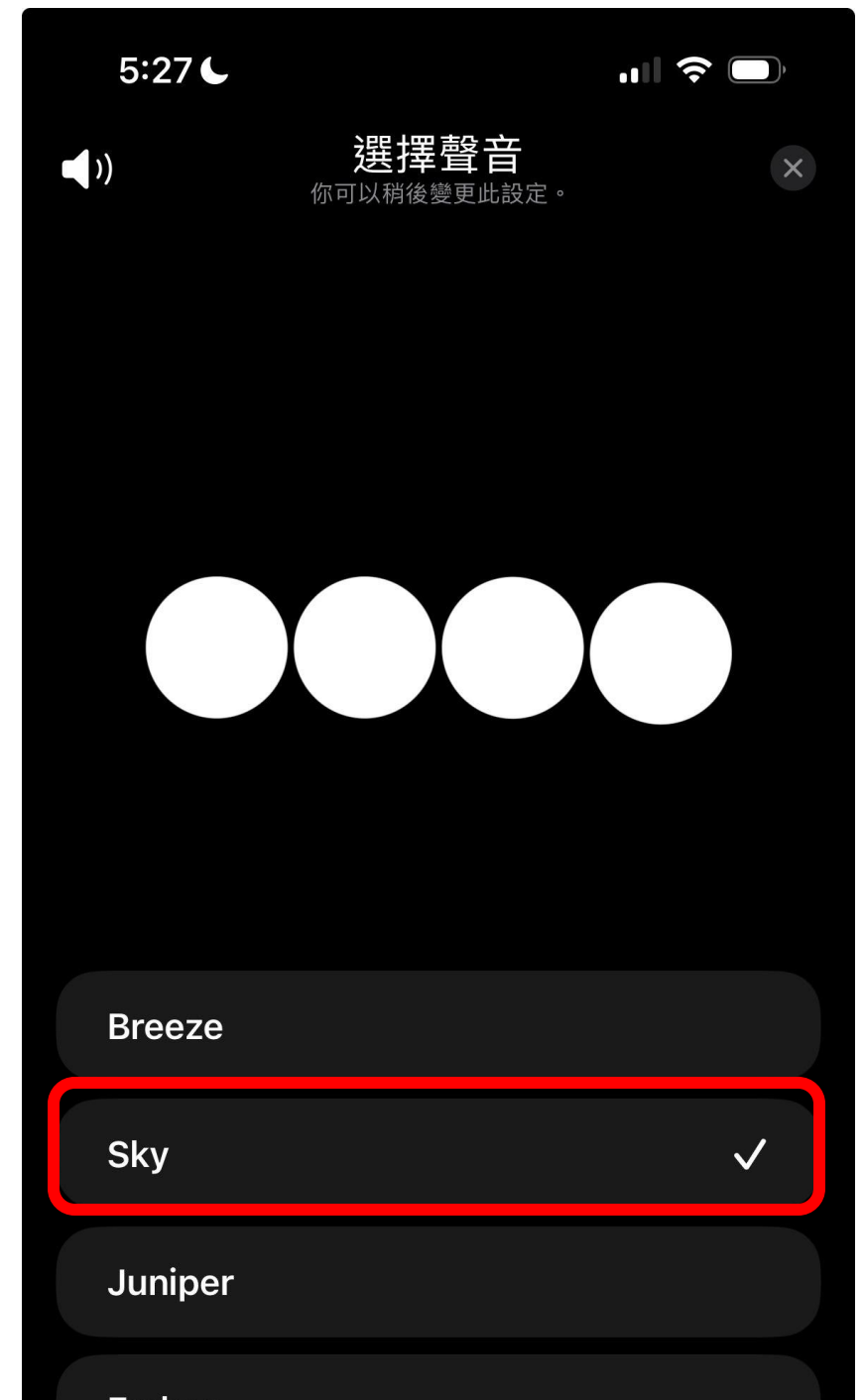
A: 快速從一數到十 Sky: 一、二、三、四 ...



會不會需要 Sky 錄很多對話啊？

- 也許不用，因為模型有 Pre-train 了？
- 也許可以用語音轉換的技術把對話中的任何人聲轉成 Sky 的聲音

<https://openai.com/index/navigating-the-challenges-and-opportunities-of-synthetic-voices/>



語音介面和文字介面的不同

The image illustrates the difference between voice and text interfaces for ChatGPT. It features two overlapping panels. The left panel shows a text input field with the prompt '說一個五千字的故事' (Tell me a 5000-character story) and the OpenAI logo. The right panel shows a voice input field with the same prompt, a microphone icon highlighted in a red square, and a response from ChatGPT: '故事的開頭：在一個遙遠的國度裡，有一個名叫伊莎貝拉的小鎮。這個小鎮坐落在群山環繞的山谷中，四季 ●' (The beginning of the story: In a distant land, there is a small town named Isabella. This town is situated in a valley surrounded by mountains, with four seasons ●).

說一個五千字的故事

說一個五千字的故事

故事的開頭：在一個遙遠的國度裡，有一個名叫伊莎貝拉的小鎮。這個小鎮坐落在群山環繞的山谷中，四季 ●

傳訊息給 ChatGPT

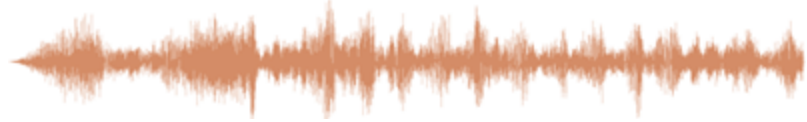
語音介面和文字介面的不同



我們來做一件有趣的事 ...

我該接「什麼事」
還是等他繼續講

語音版
語言模型



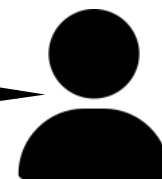
我要說個故事，山上有座廟

語音版
語言模型

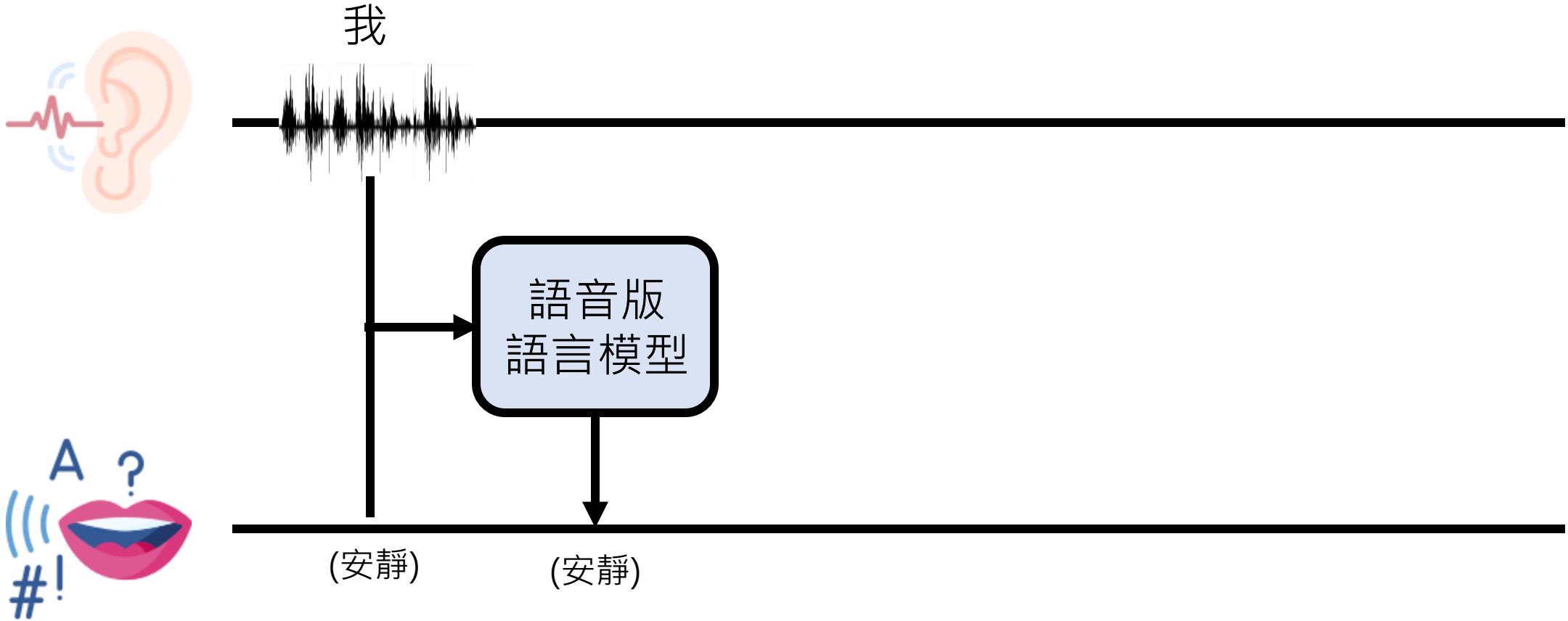
有人講話就停下來？



這不是我要聽的

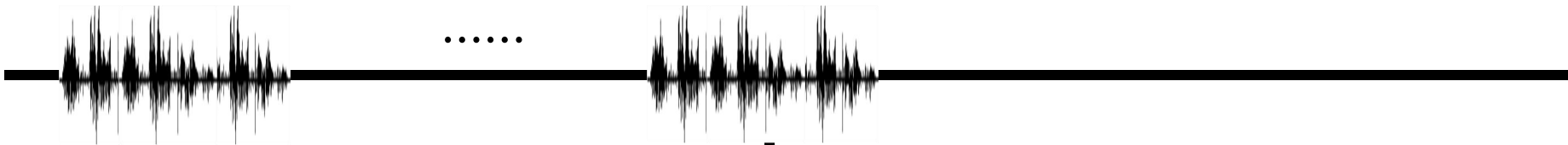


怎麼讓模型同時聽跟說

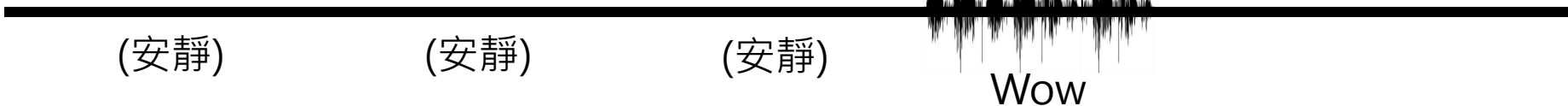


怎麼讓模型同時聽跟說

我們來做一個有趣的嘗試



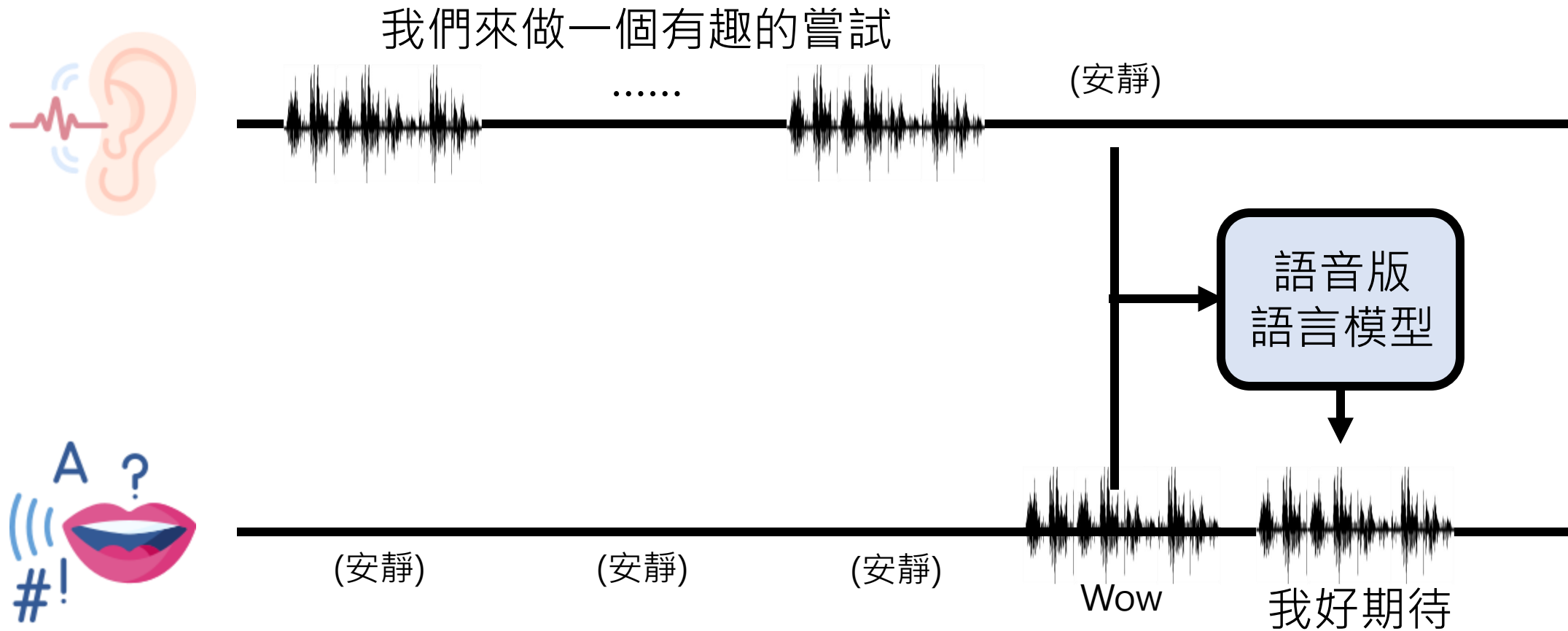
語音版
語言模型



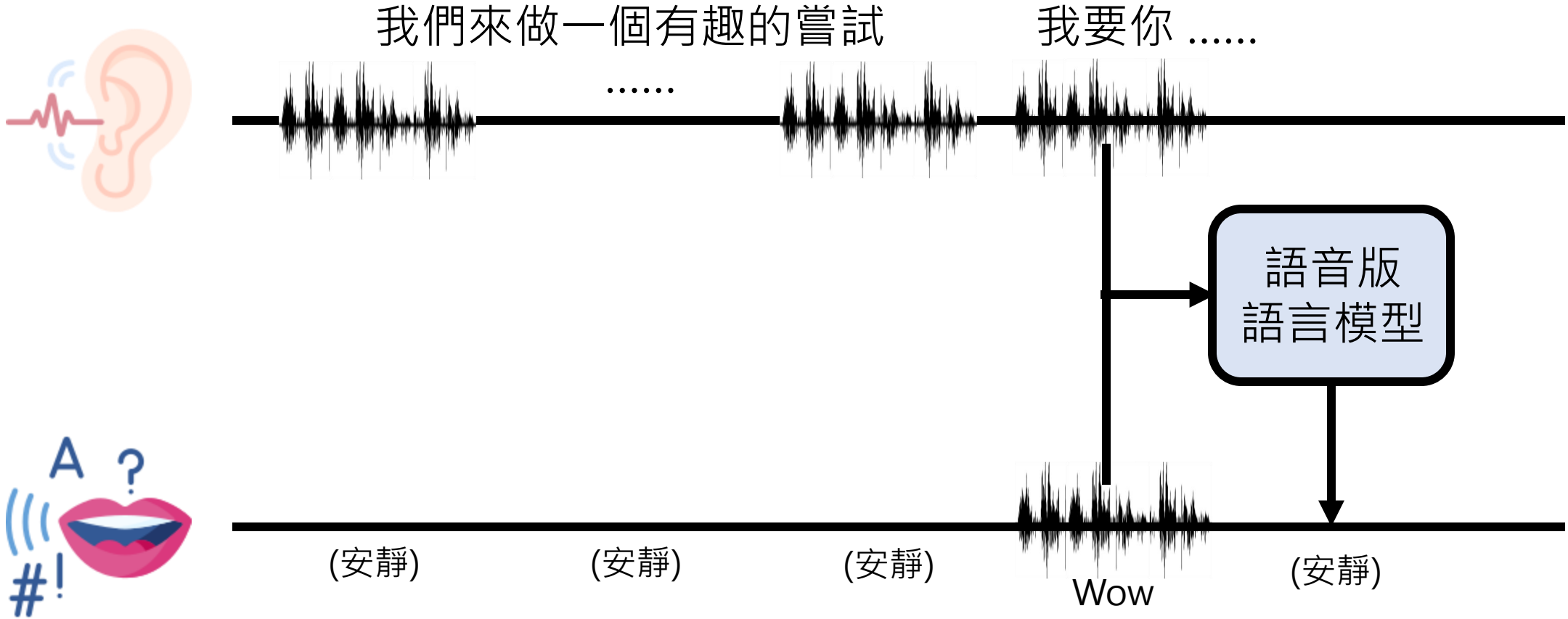
怎麼讓模型同時聽跟說

Dialogue GSLM

<https://arxiv.org/abs/2203.16502>

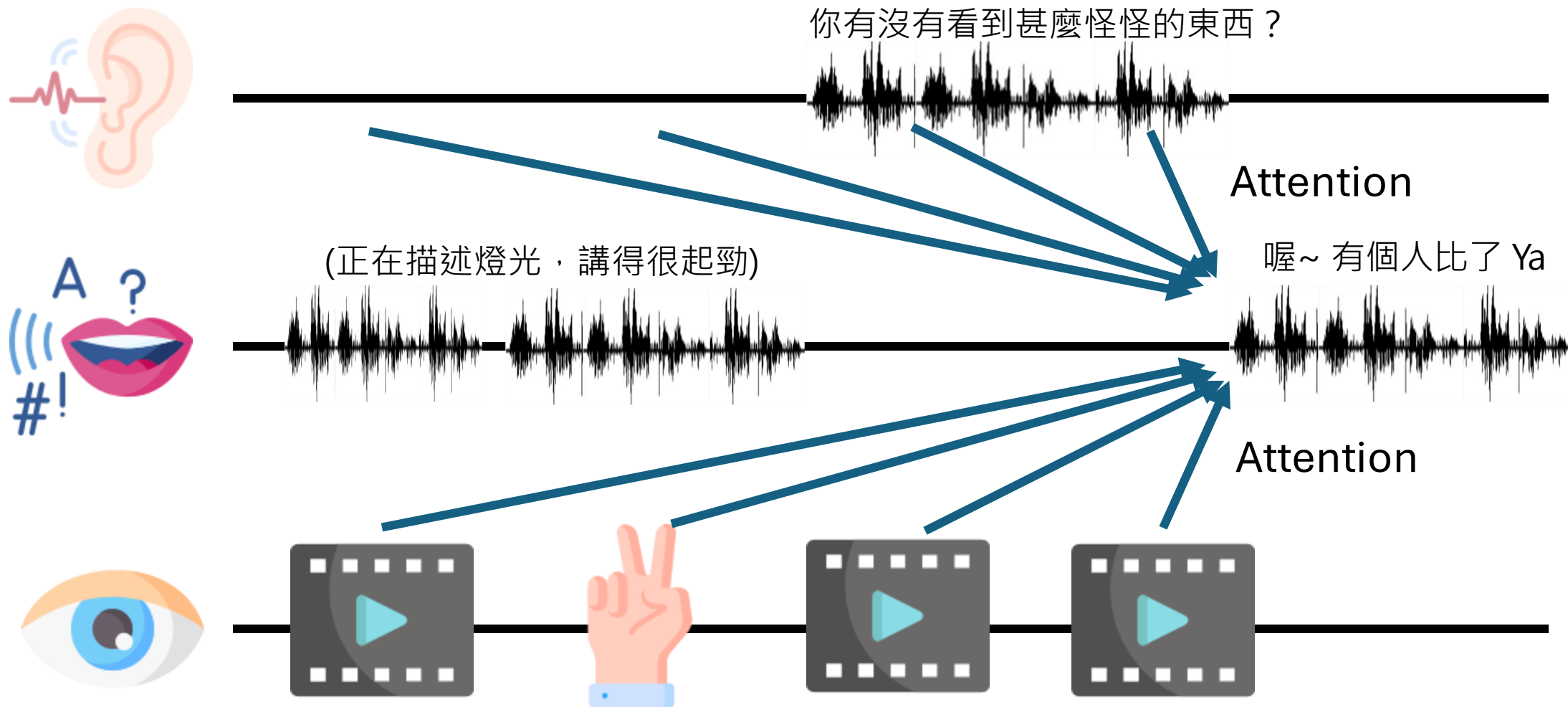


怎麼讓模型同時聽跟說



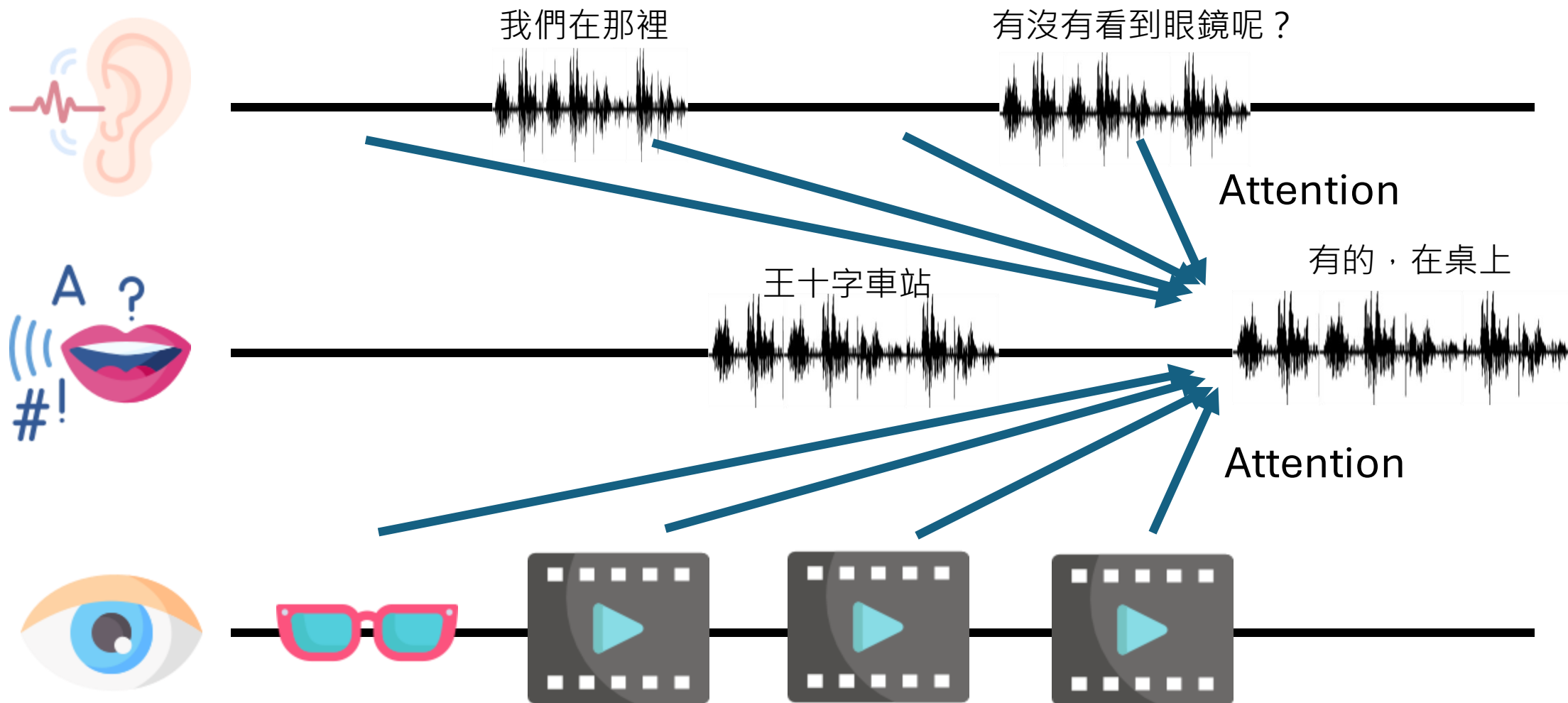
怎麼讓模型同時聽、說、看

例子來自
Open AI GPT-4o 的 demo



怎麼讓模型同時聽、說、看

例子來自
Google Project Astra Demo



更多有關語音版語言模型的論文



<https://github.com/ga642381/speech-trident>

