OpenAI GPT-4o

深度解读

ZOMI

关于本内容

1. OpenAI 春季发布：GPT-4o (o for omni)
2. 相关技术回顾：Whisper v3 – SORA – GPT4
3. GPT-4o 技术畅想：E2E 多模态大模型
4. 看大模型趋势：百模厂商的冲击 && 产业思考

OpenAI 春季发布重点发布 GPT-4o

2024.05.15

2024.05.14

Google 2024 IO大会

围绕 AI System 与 多产品亮相

多模态融合进入一个新的里程碑，丝滑

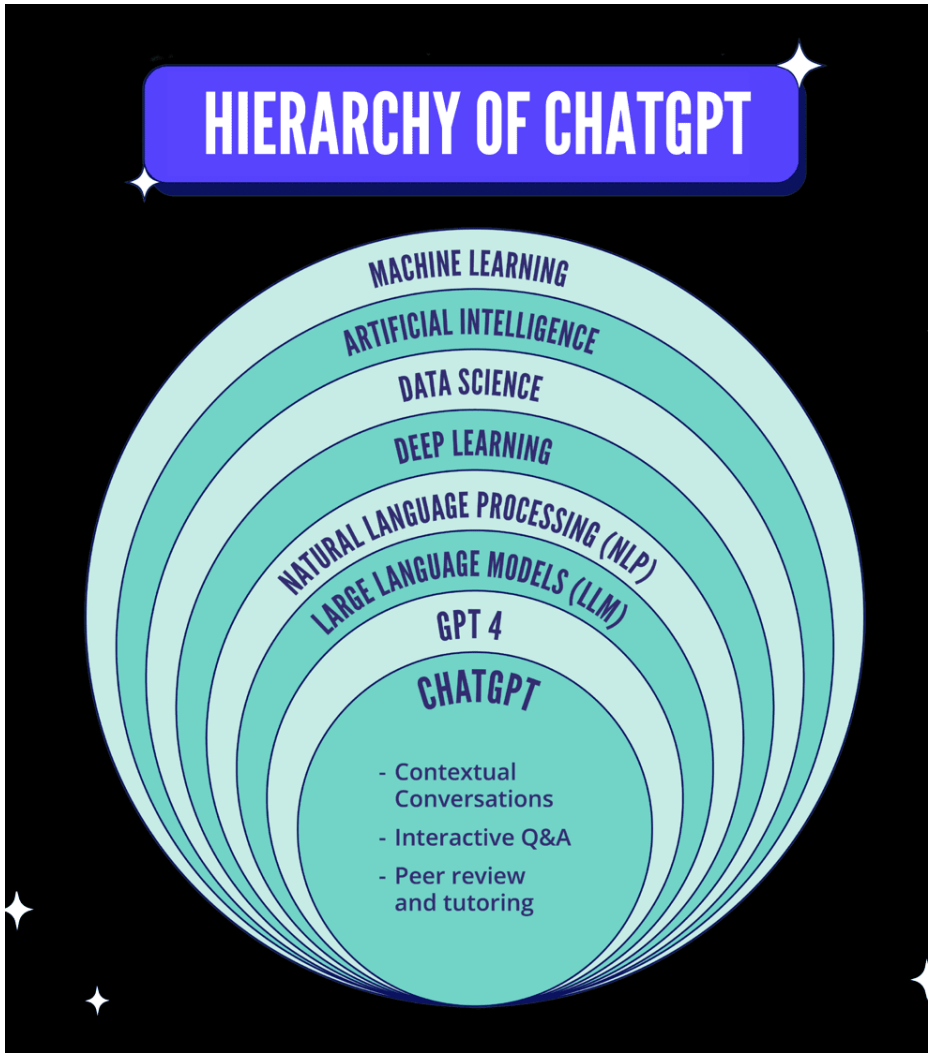
https://www.youtube.com/watch?v=-xC_dTqJLQ0&t=32s

1. OpenAI

2024 春季发布

OpenAI GPT-4o

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



May 13, 2024

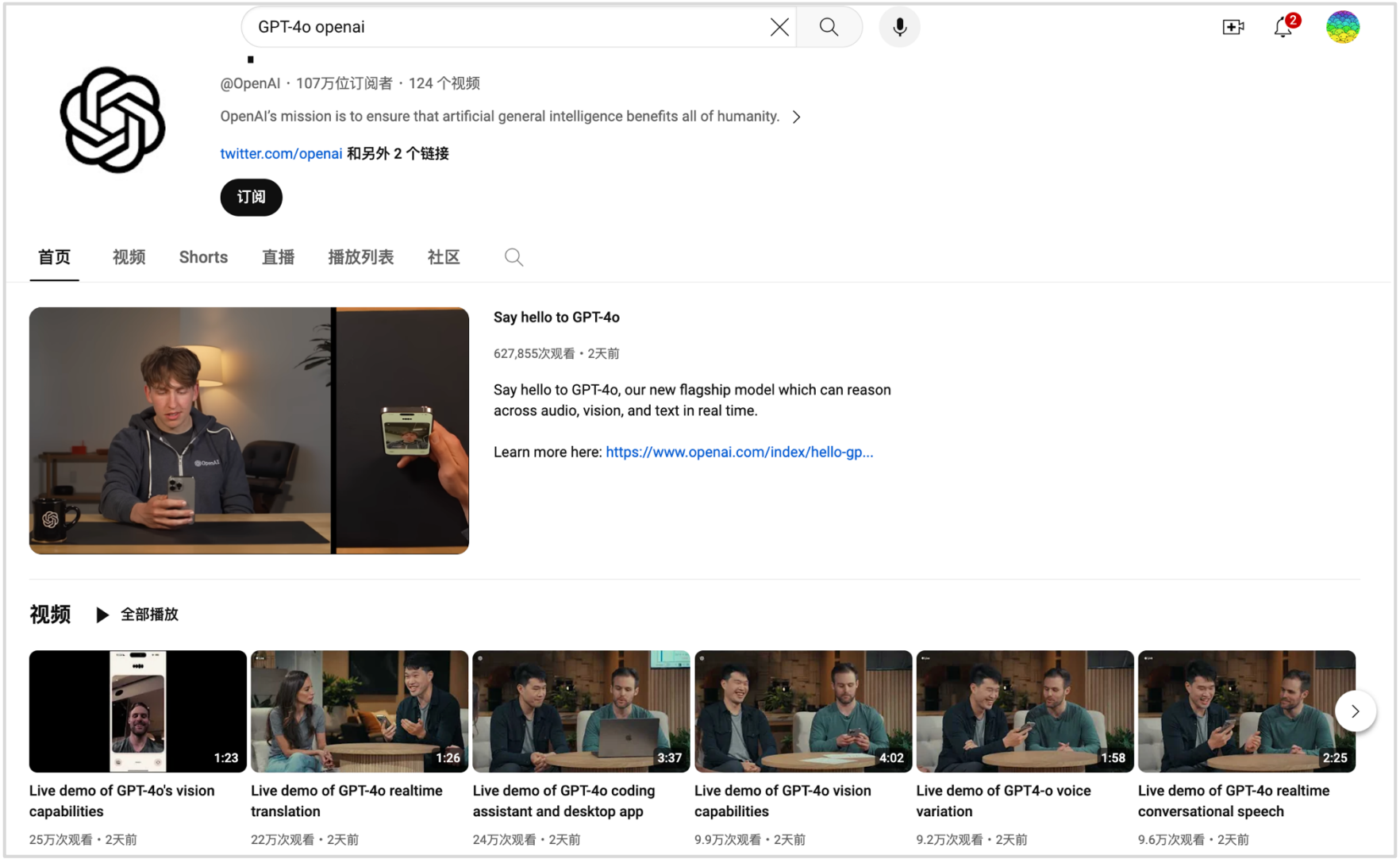
Hello GPT-4o

We're announcing GPT-4o, our new flagship model that can reason across audio, vision, and text in real time.

Contributions › Try on ChatGPT ↗ Try in Playground ↗ Rewatch live demos ›

All videos on this page are at 1x real time.

发布内容介绍


GPT-4o openai
@OpenAI · 107万位订阅者 · 124 个视频
OpenAI's mission is to ensure that artificial general intelligence benefits all of humanity.
twitter.com/openai 和另外 2 个链接
订阅
首页
视频
Shorts
直播
播放列表
社区
Say hello to GPT-4o
627,855次观看 • 2天前
Say hello to GPT-4o, our new flagship model which can reason across audio, vision, and text in real time.
Learn more here: https://www.openai.com/index/hello-gp...
视频
全部播放
1:23
Live demo of GPT-4o's vision capabilities
25万次观看 • 2天前
1:26
Live demo of GPT-4o realtime translation
22万次观看 • 2天前
3:37
Live demo of GPT-4o coding assistant and desktop app
24万次观看 • 2天前
4:02
Live demo of GPT-4o vision capabilities
9.9万次观看 • 2天前
1:58
Live demo of GPT4-o voice variation
9.2万次观看 • 2天前
2:25
Live demo of GPT-4o realtime conversational speech
9.6万次观看 • 2天前

GPT-4o contributions 看组织贡献

- Pre-training leads
- Post-training leads
- Architecture leads
- Optimization leads
- Long-context lead
- Pre-training Data leads
- Tokenizer lead
- Human data leads
- Eval lead
- Data flywheel lead
- Inference lead
- …

Language Large Model

- Multimodal lead
- Post-Training Multimodal lead
- Audio Pre-Training leads
- Audio Post-Training leads
- Visual perception leads
- Visual generation leads
- Data acquisition leads
- Data infrastructure leads
- Human data lead
- Encoders leads
- Decoders leads
- …

Multimodal Large Model

GPT-4o contributions 看组织贡献

- **核心组织：**语言大模型LLM 项目 + 多模态大模型MLM项目 + 平台项目 + 启动和部署项目；
- **组成：**13个小组，产品相关人 400+；

- **语言大模型LLM 项目**：
  - 16个小组，220人+；
  - 长文本、预训练、数据飞轮、Tokenizer；

- **多模态大模型MLM项目：**
  - 20个小组，106人+；
  - 语音预训练、视觉感知与生成、编解码；

- **平台项目：**
  - 11个小组，58人+；
  - 集群管理、模型发布、数据工程；

- **模型启动与部署：**
  - 25人+；
  - 安全隐私、上线推理、部署；

相关项目

Our research

Overview

Index

Latest advancements

GPT-4o

DALL·E 3

Sora

Hello GPT-4o

Our research

Overview

Index

Latest advancements

GPT-4

DALL·E 3

Sora

ChatGPT

For Everyone

For Teams

For Enterprises

ChatGPT login ↗

API

Platform overview

Pricing

Safety overview

Safety overview

Safety standards

Teams

Safety Systems

Preparedness

Superalignment

Company

About us

News

Our Charter

Security

Residency

Careers

Terms & policies

Terms of use

Privacy policy

Brand guidelines

Other policies

3. 相关技术回顾

1 Whisper v3

https://github.com/openai/whisper

Multitask training data (680k hours)

English transcription

- "Ask not what your country can do for ⋯"
- Ask not what your country can do for ⋯

Any-to-English speech translation

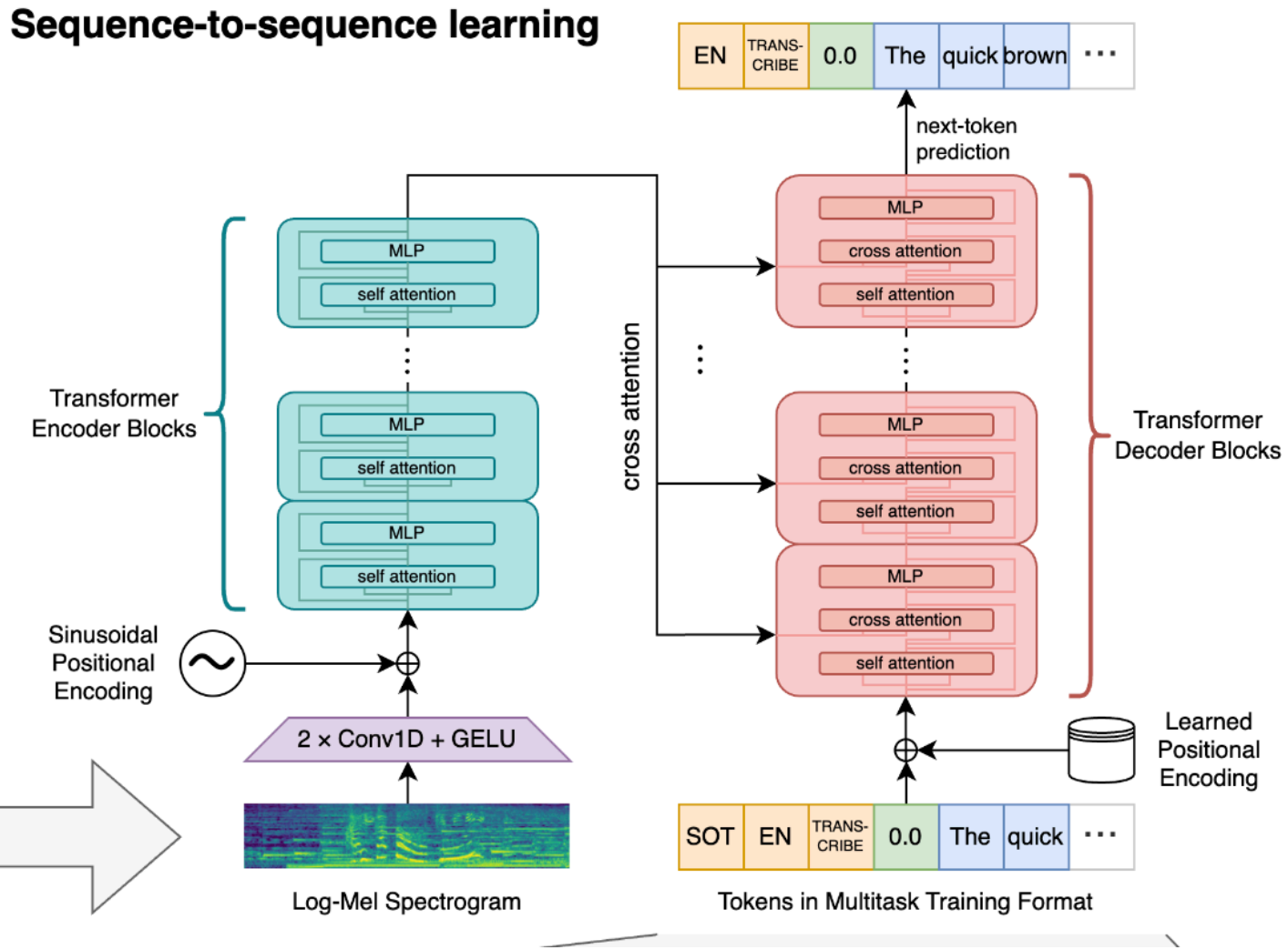
- "El rápido zorro marrón salta sobre ⋯"
- The quick brown fox jumps over ⋯

Non-English transcription

- "언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯"
- 언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯

No speech

- (background music playing)
- ∅



1 Whisper v3

Robust Speech Recognition via Large-Scale Weak Supervision

Alec Radford [*1] **Jong Wook Kim** [*1] **Tao Xu** [1] **Greg Brockman** [1] **Christine McLeavey** [1] **Ilya Sutskever** [1]

Abstract

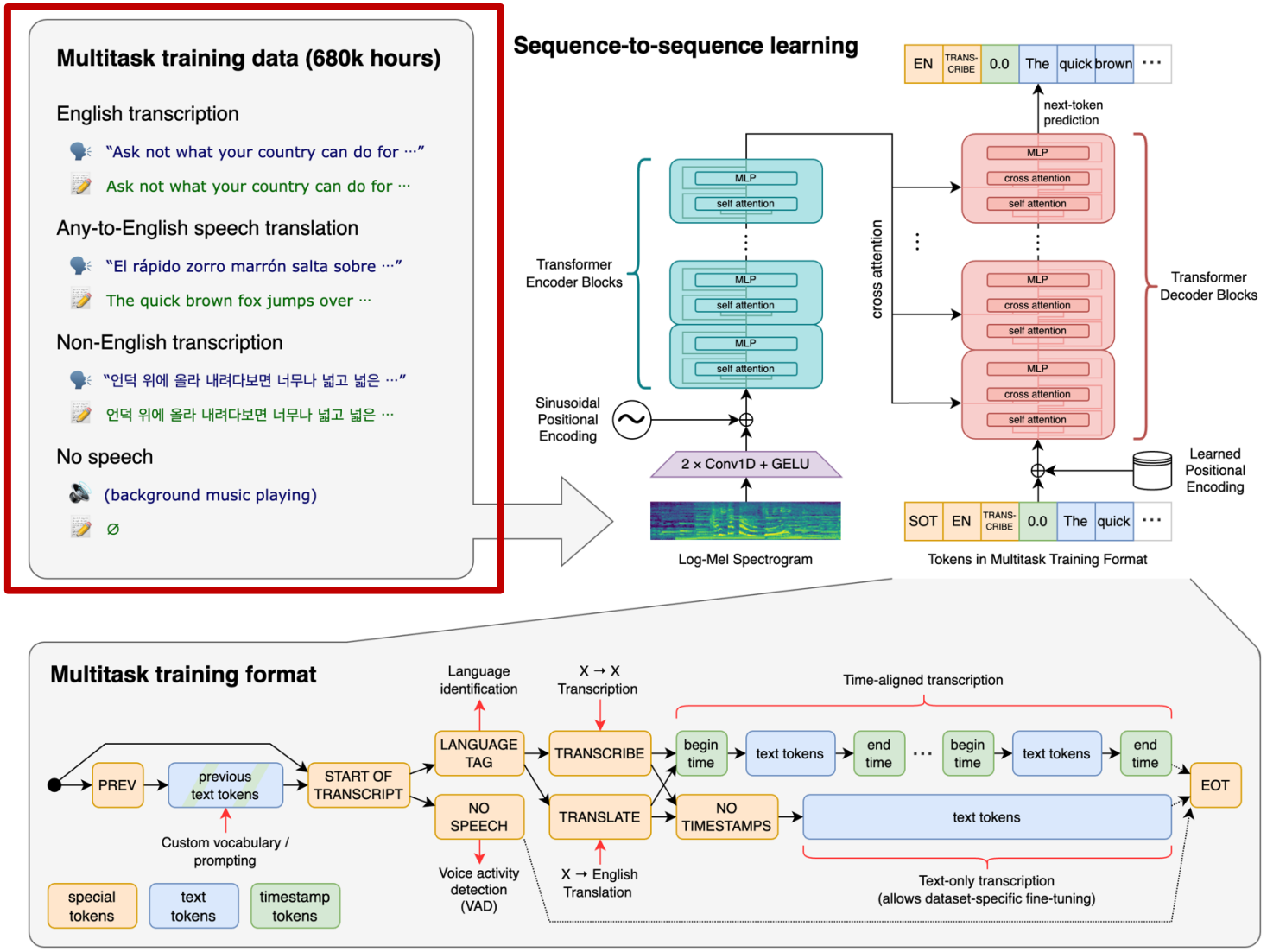
We study the capabilities of speech processing systems trained simply to predict large amounts of transcripts of audio on the internet. When scaled to 680,000 hours of multilingual and multitask supervision, the resulting models generalize well to standard benchmarks and are often competitive with prior fully supervised results but in a zero-shot transfer setting without the need for any fine-tuning. When compared to humans, the models approach their accuracy and robustness. We are releasing models and inference code to serve as a foundation for further work on robust speech processing.

methods are exceedingly adept at finding patterns within a training dataset which boost performance on held-out data from the same dataset. However, some of these patterns are brittle and spurious and don't generalize to other datasets and distributions. In a particularly disturbing example, Radford et al. (2021) documented a 9.2% increase in object classification accuracy when fine-tuning a computer vision model on the ImageNet dataset (Russakovsky et al., 2015) without observing any improvement in average accuracy when classifying the same objects on seven other natural image datasets. A model that achieves "superhuman" performance when trained on a dataset can still make many basic errors when evaluated on another, possibly precisely because it is exploiting those dataset-specific quirks that humans are oblivious to (Geirhos et al., 2020).

1 Whisper 作用

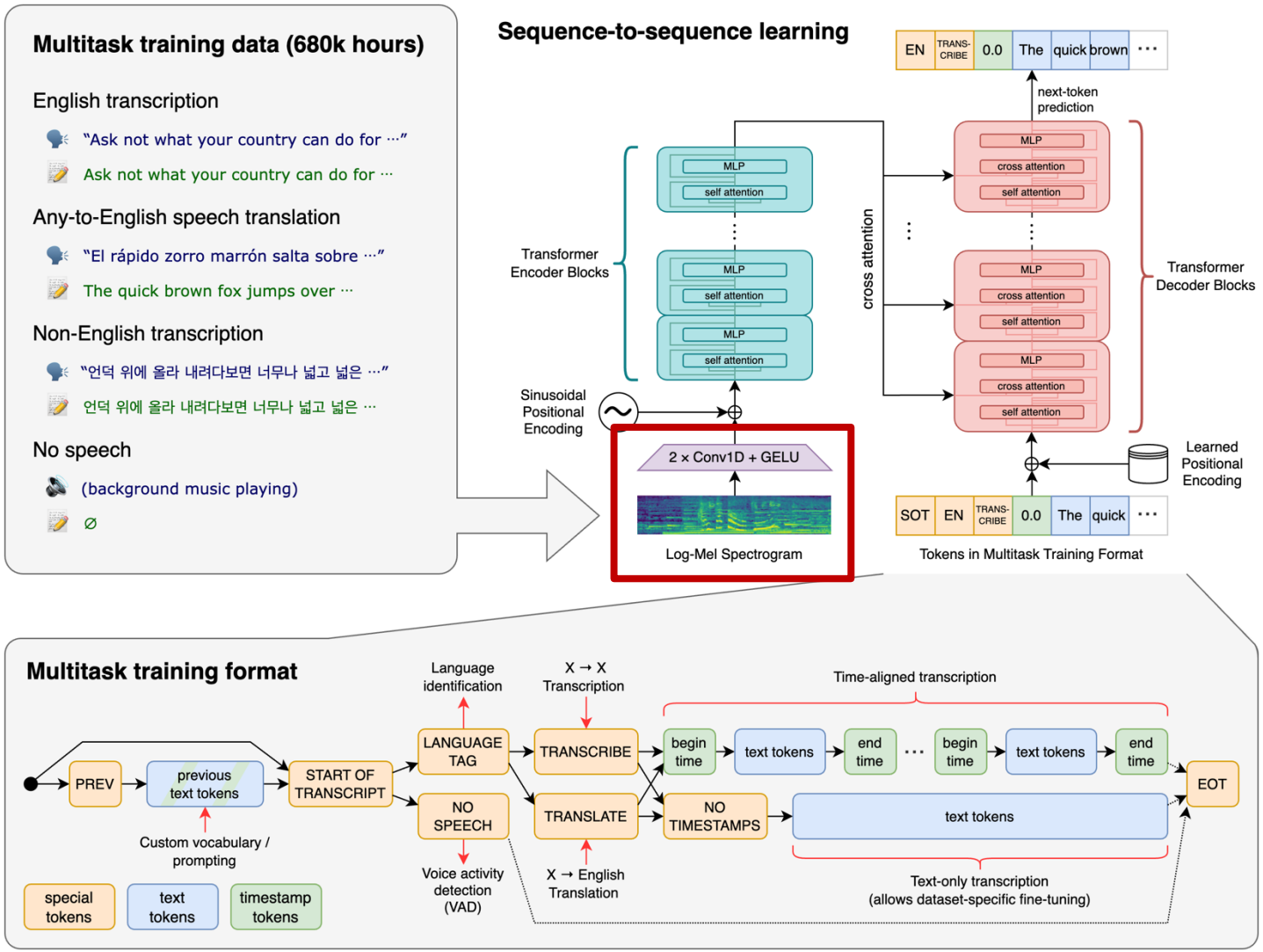
```
1
00:00:00,000 --> 00:00:07,775
嗨 大家好 我是 ZOMI

2
00:00:07,775 --> 00:00:10,300
今天来到分布式训练系列里面的

3
00:00:10,300 --> 00:00:11,520
大模型算法结构

4
00:00:11,520 --> 00:00:14,150
那聊到大模型的算法结构

5
00:00:14,150 --> 00:00:15,850
主要是去看看

6
00:00:15,850 --> 00:00:18,400
大模型算法的一个整体的发展

7
00:00:18,400 --> 00:00:20,975
从没有到有，从小到大
```

1 Whisper 原理



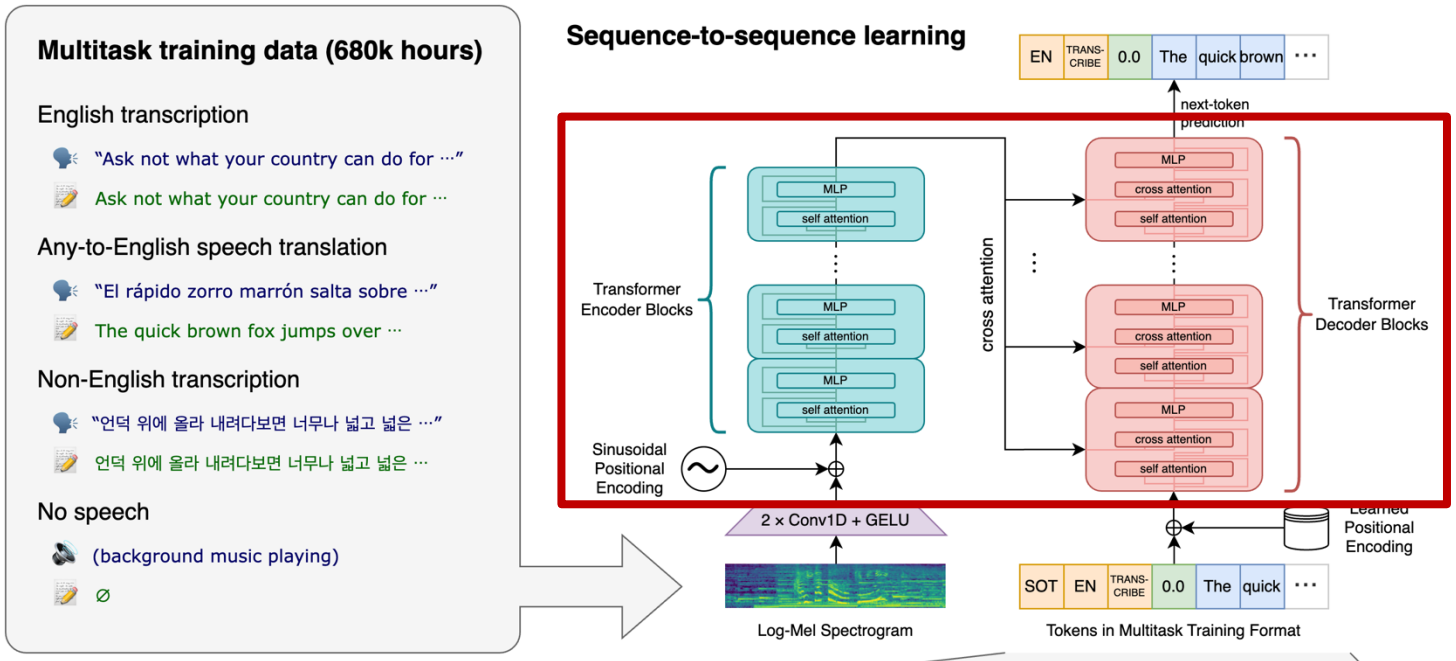
- Whisper 使用 680k 小时的音频数据训练（这些数据包括各种语言、口音、说话速度和背景噪声等），通过学习音频信号中的特征和模式，最终实现对语音的准确识别。
- 先将信号转换为频谱图，然后使用 Mel 频率倒谱系数（MFCC）特征提取法，将频谱图转换为特征向量。

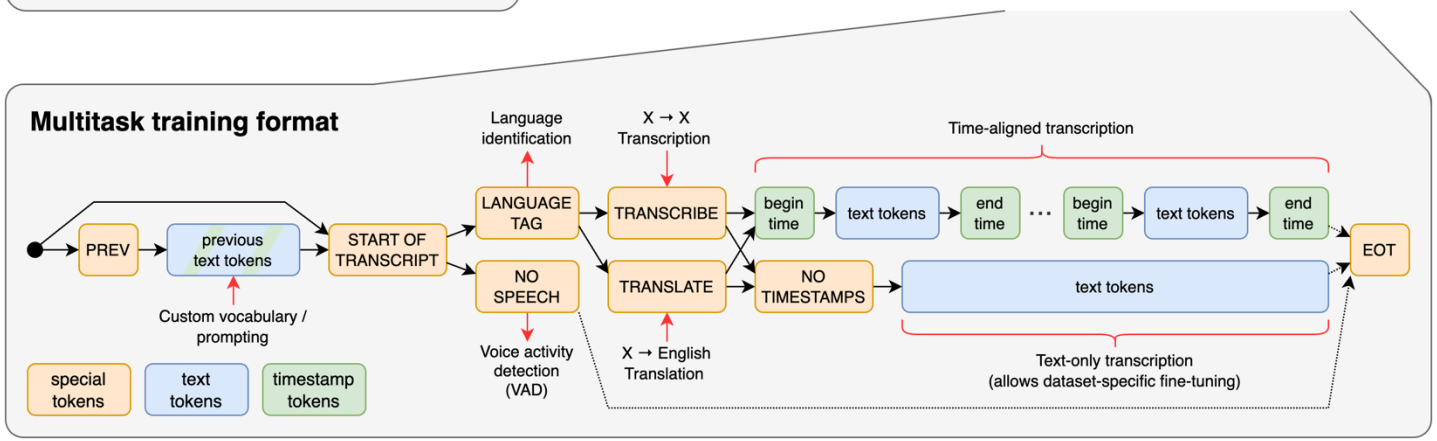
1 Whisper 原理



- Whisper 将特征向量作为输入，将其与对应文本标签进行匹配。通过反向传播算法，不断调整神经网络的权重和偏置，使得模型能够更准确地预测语音对应的文本。
- 使用 CTC（Connectionist Temporal Classification）解码算法，将输出概率分布映射到最可能的文本序列。

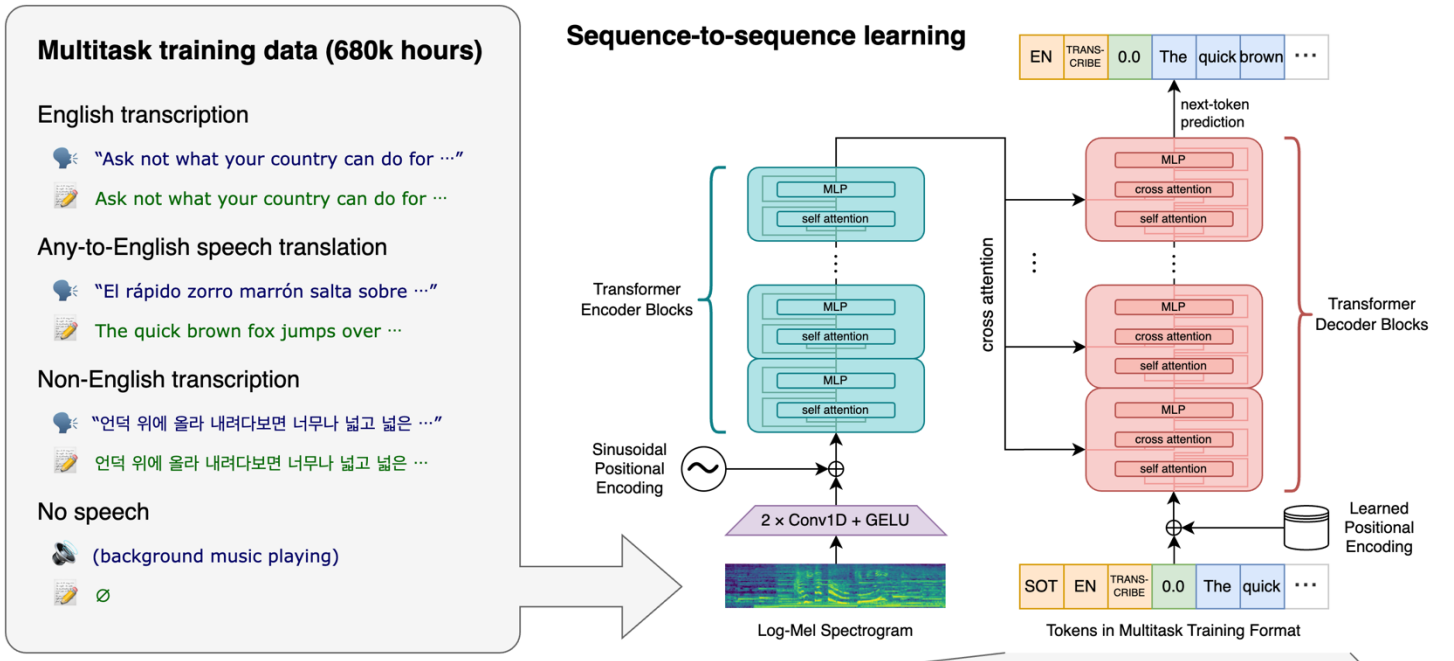
1 Whisper 原理

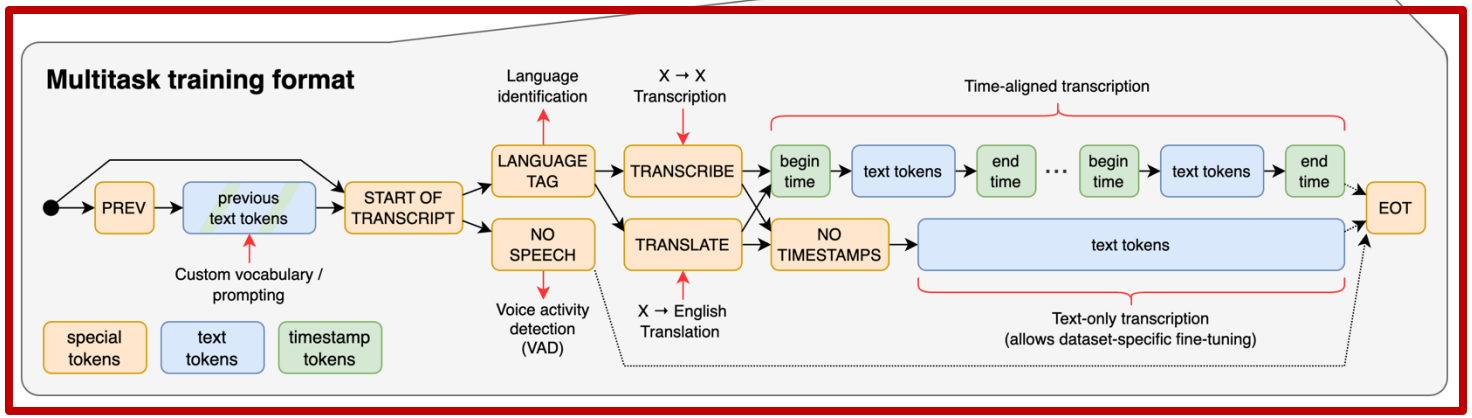




- Whisper 使用经典 Seq2Seq，主要原理就是通过两个 Transformer 做节编解码器；
- 编码器任务就是把机器翻译中源语言的句子给编码成向量；
- 解码器的作用就是把编码器编码好的向量再通过 Transformer 一点点生成语言的句子，其中解码器的作用类似于文本生成。

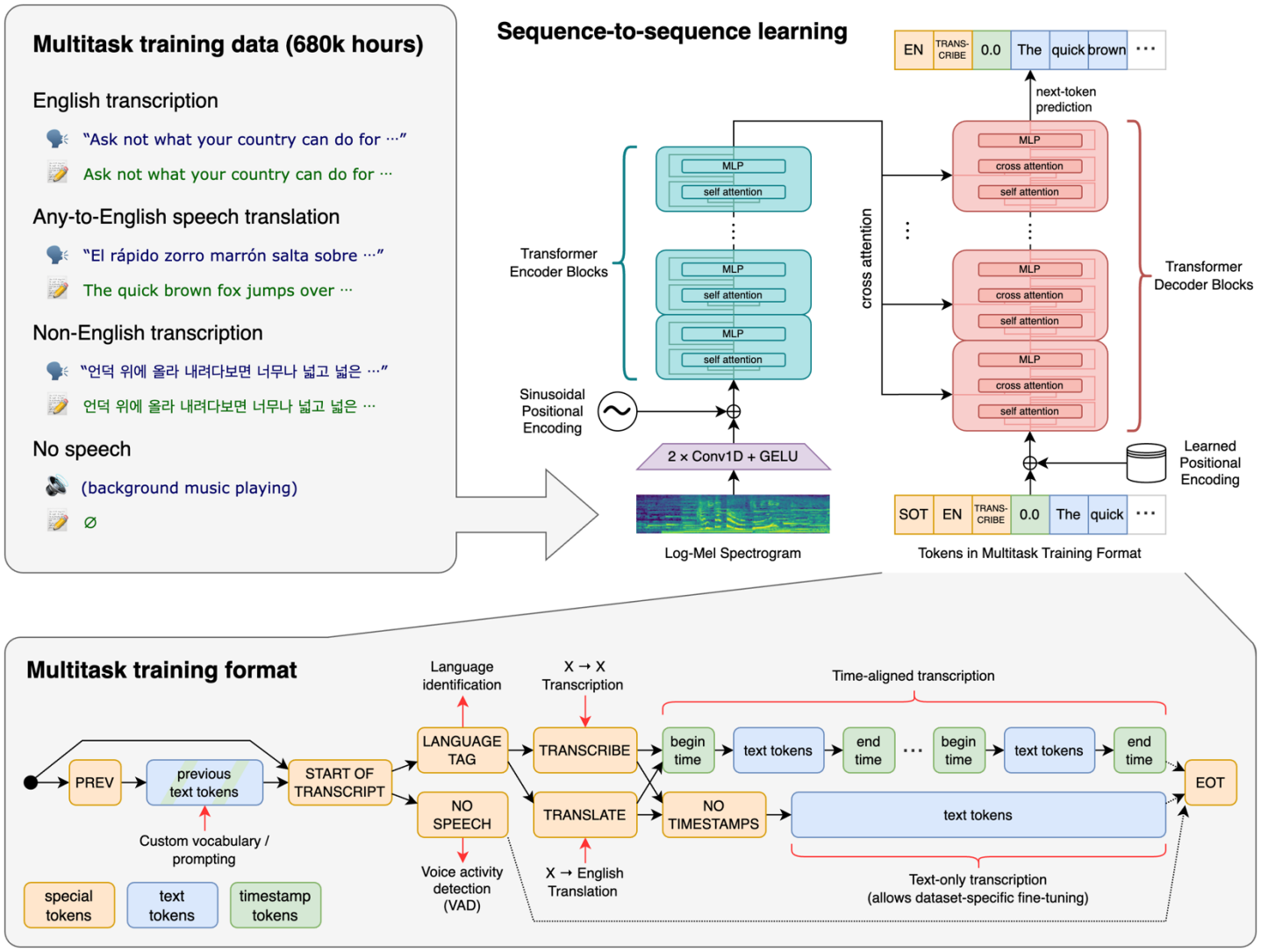
1 Whisper 原理





- 不同任务由解码器预测 token 序列表示，从而使得一个模型能够处理多个任务。
- START OF TRANSCRIPT token 后，当前无人说话则识别为 NO SPEECH。有人说话则识别出当前语音所属语言 LANGUAGE TAG。
- 最后又分为带时间戳和不带时间戳。最后到达 EOT token，整个流程结束。

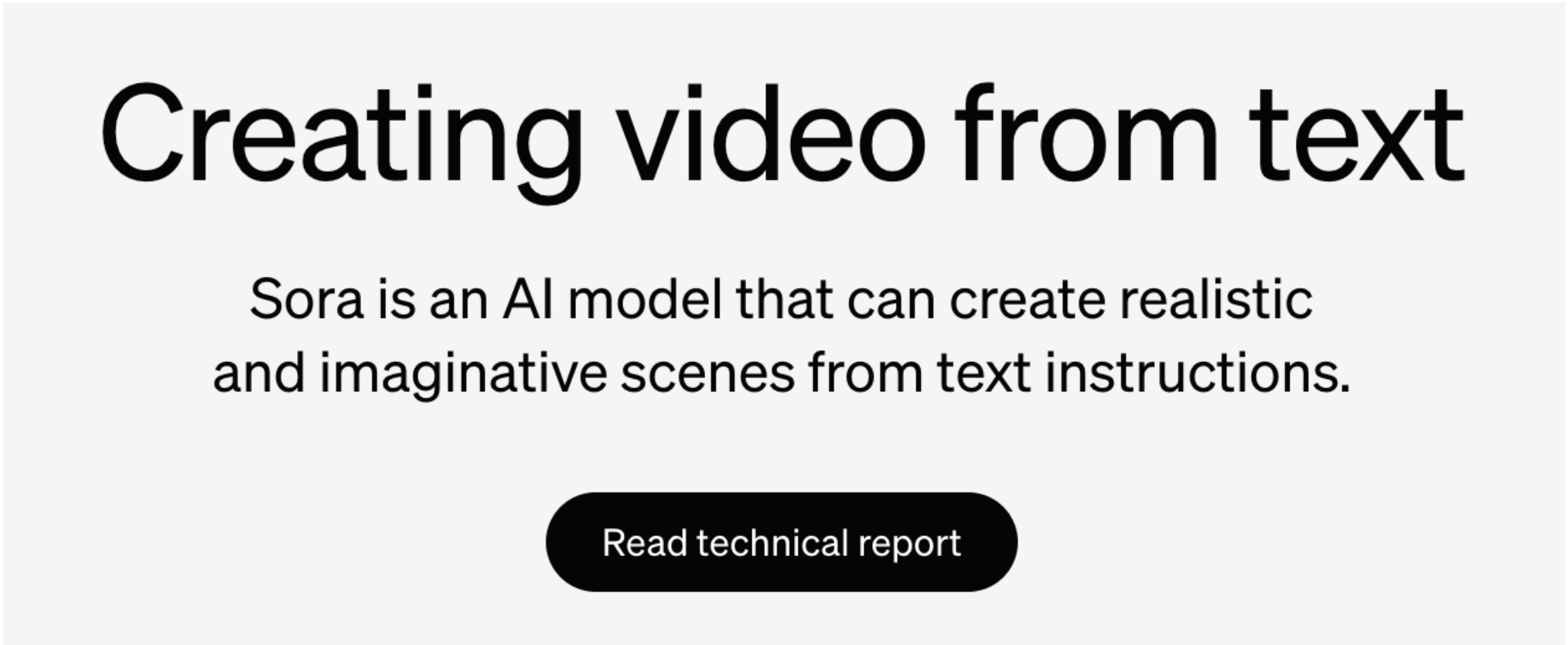
1 Whisper 作用



- 除了可以用于语音识别，Whisper 还能包括多语种语音识别、语音翻译、口语语言识别以及语音活动检测。
- 从而使得单个模型可以替代传统语音处理管道中的多个阶段。多任务训练格式使用一系列特殊的符号作为任务指示符或分类目标。

2 SORA 文生视频出现

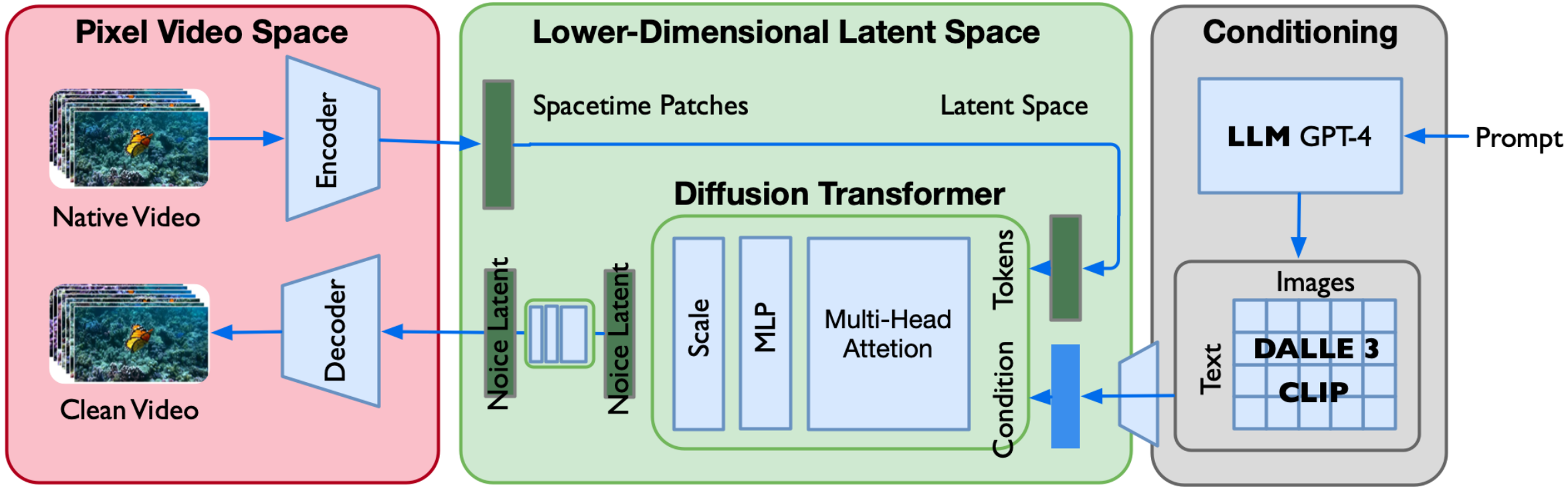
https://openai.com/index/sora/



2 SORA 模型结构

- SORA 模型结构可以表示：

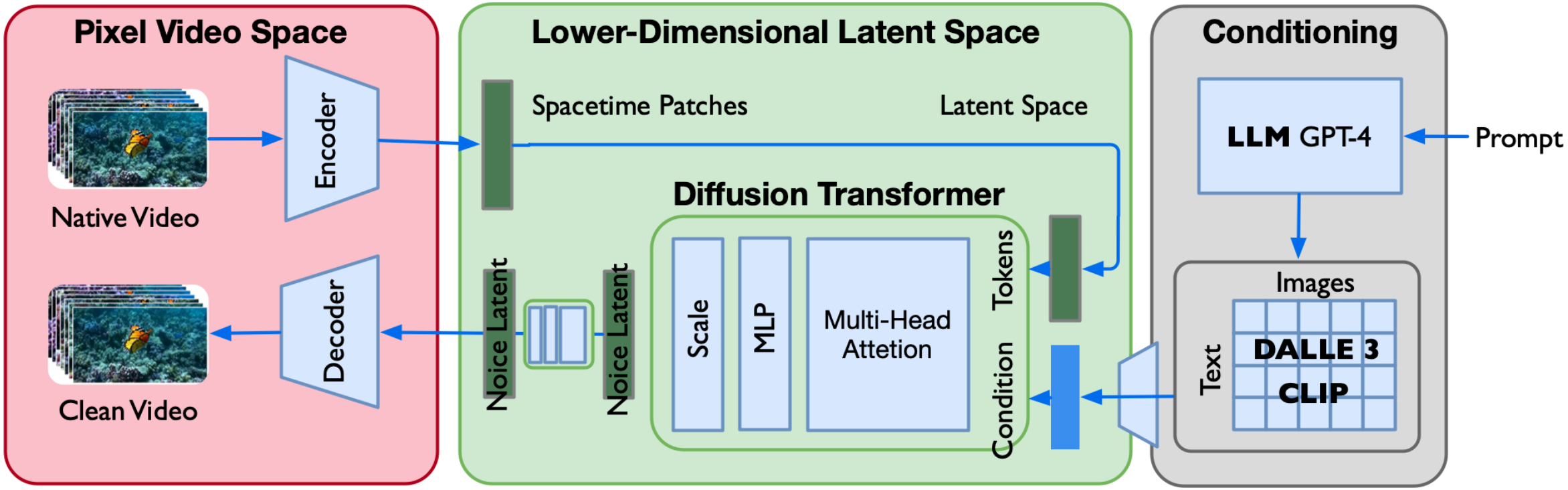
$$\text{SORA} = [\text{VAE encoder} + \text{DiT (DDPM)} + \text{VAE decoder} + \text{CLIP}]$$



2 SORA 模型结构

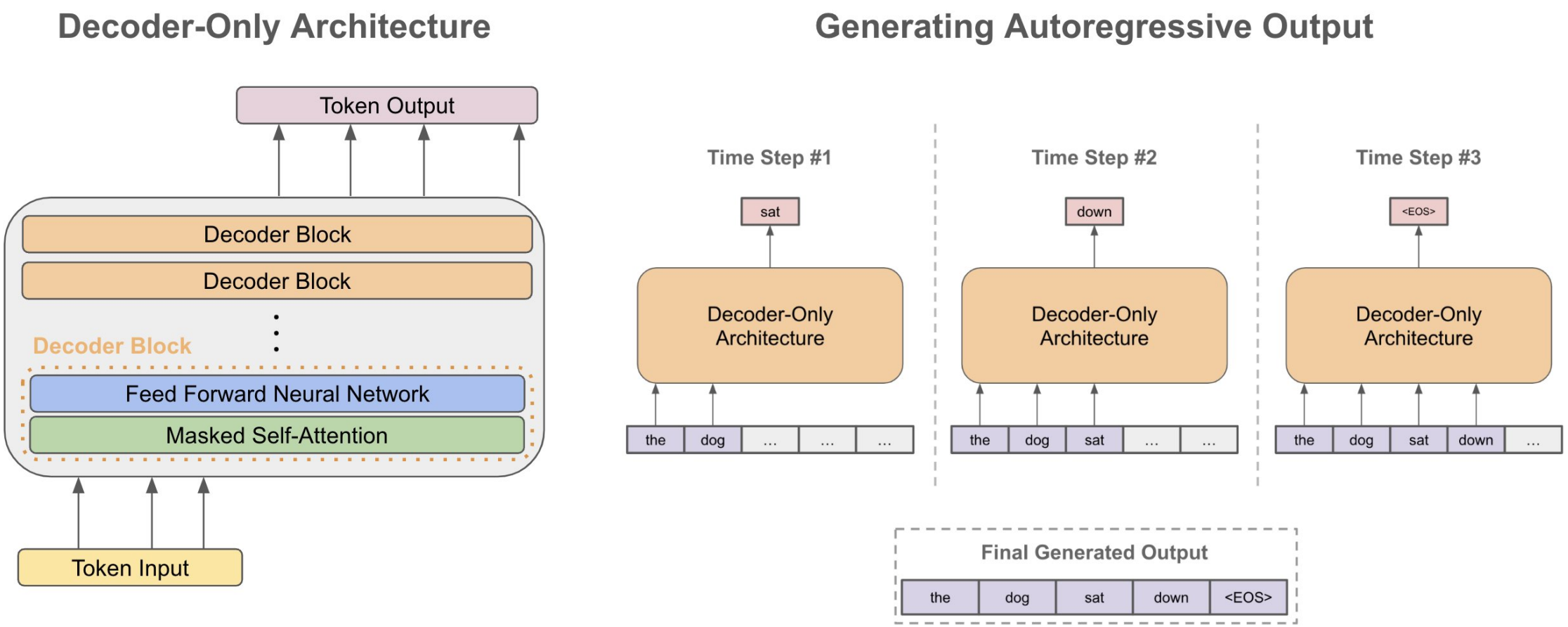
- SORA 模型结构可以表示：

$$\text{SORA} = [\text{VAE encoder} + \text{DiT (DDPM)} + \text{VAE decoder} + \text{CLIP}]$$



2 SORA 深度剖析

3 GPT 3 模型结构

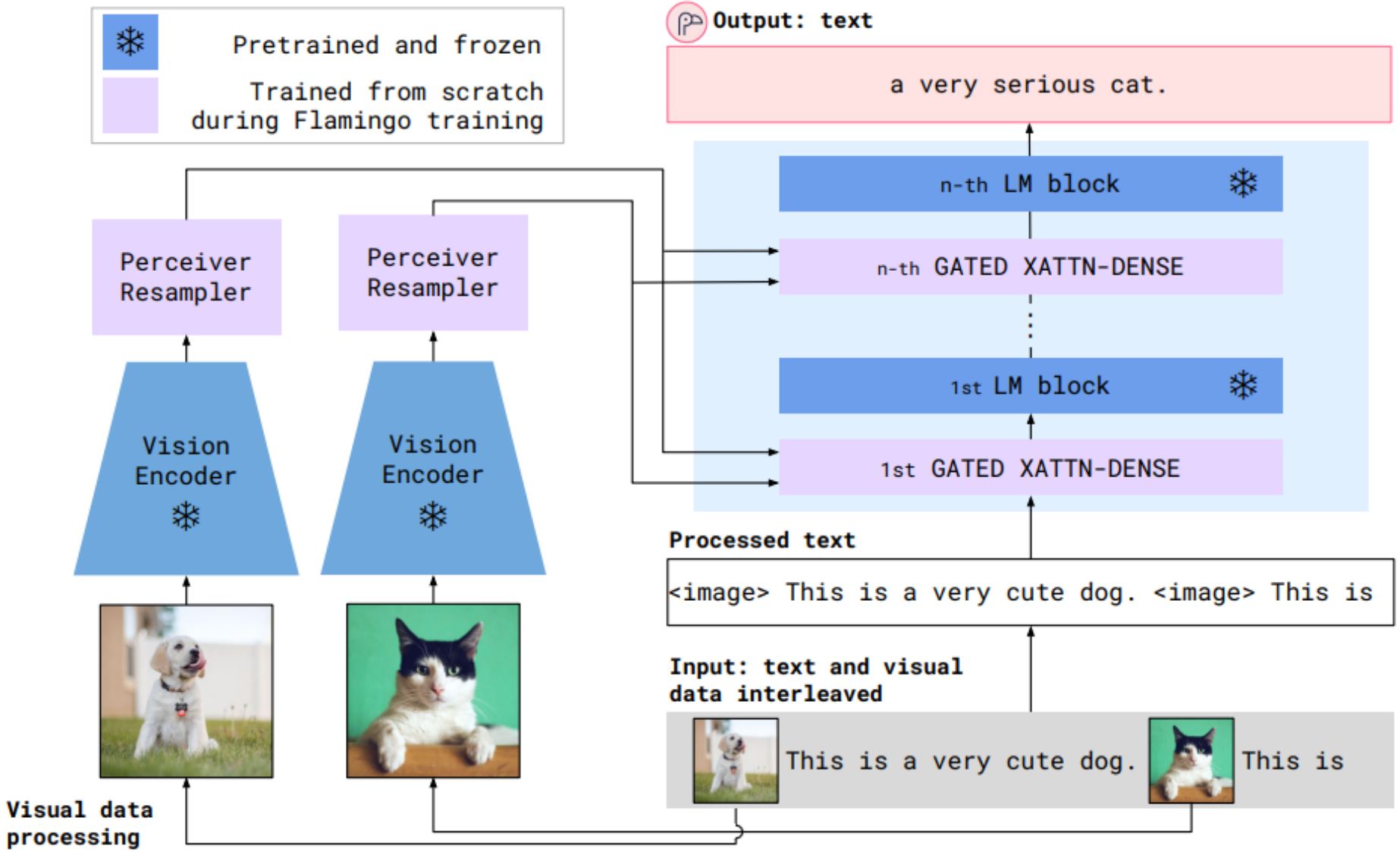

Decoder-Only Architecture
Token Output
Decoder Block
Decoder Block
Decoder Block
Feed Forward Neural Network
Masked Self-Attention
Token Input
Generating Autoregressive Output
Time Step #1
Time Step #2
Time Step #3
sat
down
<EOS>
Decoder-Only Architecture
Decoder-Only Architecture
Decoder-Only Architecture
the
dog
...
...
...
the
dog
sat
...
...
the
dog
sat
down
...
Final Generated Output
the
dog
sat
down
<EOS>

3 GPT 4 参数量增加

	Parameters	Decoder layers	Context lenght	Hidden layer size
GPT-1	117 million	12	512	768
GPT-2	1.5 billion	48	1024	1600
GPT-3	175 billion	96	2048	12288
GPT-4	1.76 trillion	120	8000*	20k*

*Data subject to confirmation by OpenAI. Last updated: July 2023.

3 GPT 4 模型结构猜测



4 AI 对话、AI 助手

Whisper

"Turn on the party light and play cool music at 8 pm every other Tuesday"

Turn on the party light and play cool music at 8 pm every other Tuesday

ChatGPT

```
if weekday() == Tue:
  …
  … <SMART HOME API CODE>
return <confirmation msg>
```

"Sure Jim, I have set it up according to your preference"

+ 3 second voice sample

VALL-E

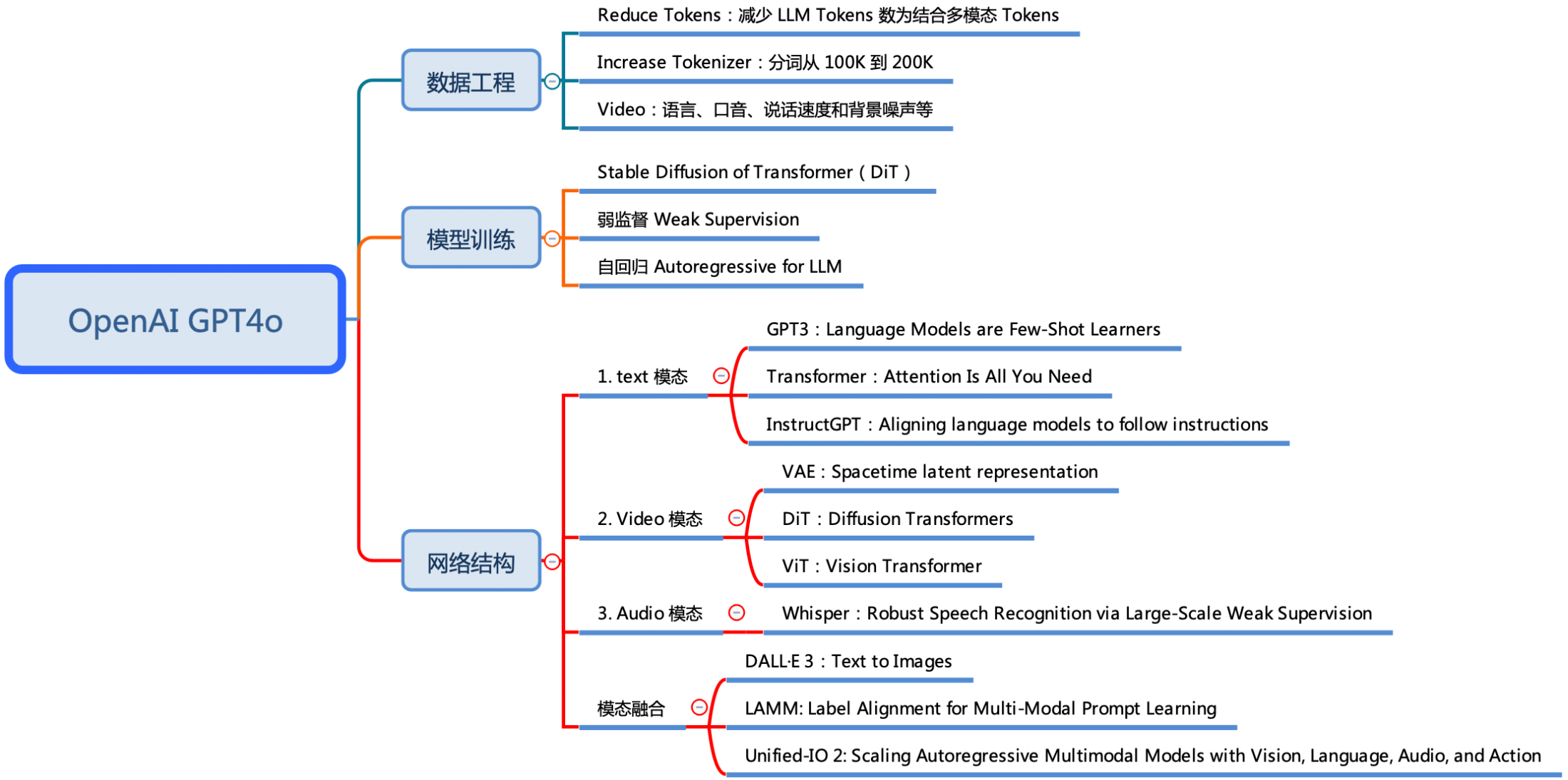
3. GPT-4o

技术原理（畅想）

米国五星上将麦克阿瑟：

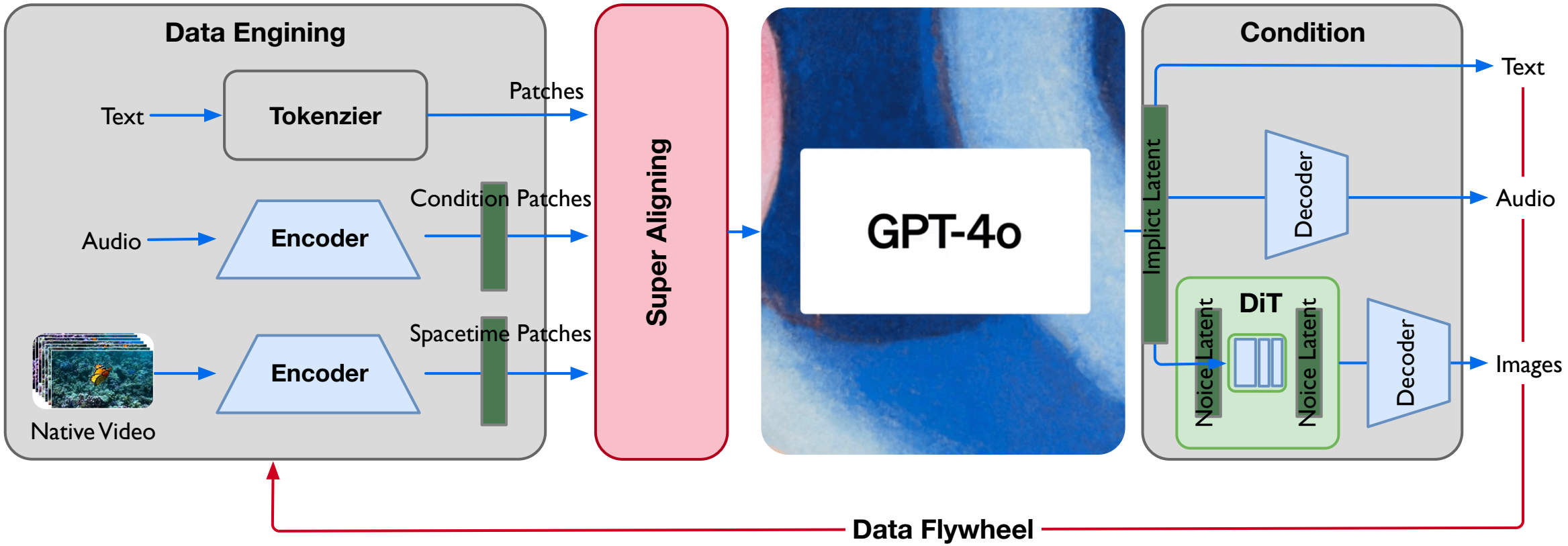
牛逼不是一天练成的

思维导图


OpenAI GPT4o
数据工程
Reduce Tokens：减少 LLM Tokens 数为结合多模态 Tokens
Increase Tokenizer：分词从 100K 到 200K
Video：语言、口音、说话速度和背景噪声等
模型训练
Stable Diffusion of Transformer（DiT）
弱监督 Weak Supervision
自回归 Autoregressive for LLM
网络结构
1. text 模态
GPT3：Language Models are Few-Shot Learners
Transformer：Attention Is All You Need
InstructGPT：Aligning language models to follow instructions
2. Video 模态
VAE：Spacetime latent representation
DiT：Diffusion Transformers
ViT：Vision Transformer
3. Audio 模态
Whisper：Robust Speech Recognition via Large-Scale Weak Supervision
模态融合
DALL·E 3：Text to Images
LAMM: Label Alignment for Multi-Modal Prompt Learning
Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action

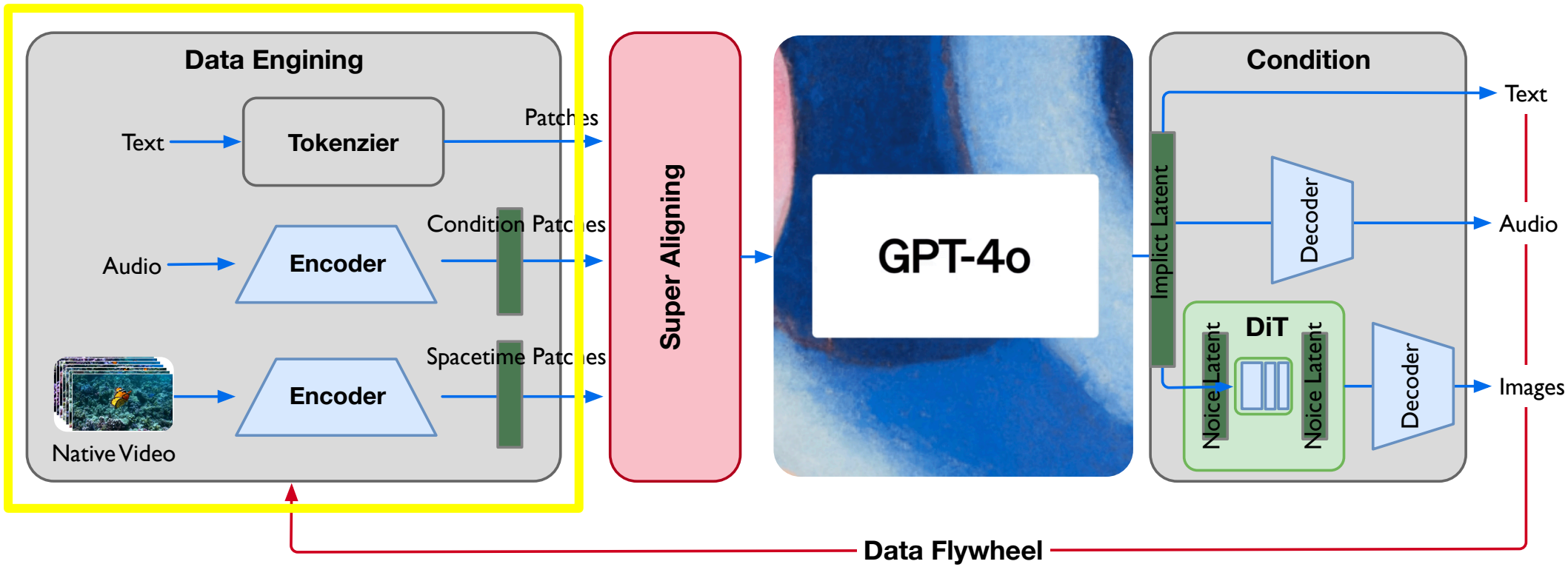
GPT-4o 原理解读

$$\text{GPT} - 4\text{o} = \text{DataEngineering} + \text{SuperAligning} + [\text{Transformer Decoder}] \times l + \text{Condition}$$



GPT-4o 原理解读：Data Engineering

- LLM 仍然是主战场，投入人力超1/2，新的 Tokenizer 新的 Vocab Size；
- Audio 参考 Whisper v3 与 Text 结合作为 Multitask training format 再编码；
- Video/Image 借鉴 Sora 的 Spacetime Patches 极致编码压缩；



数据工程

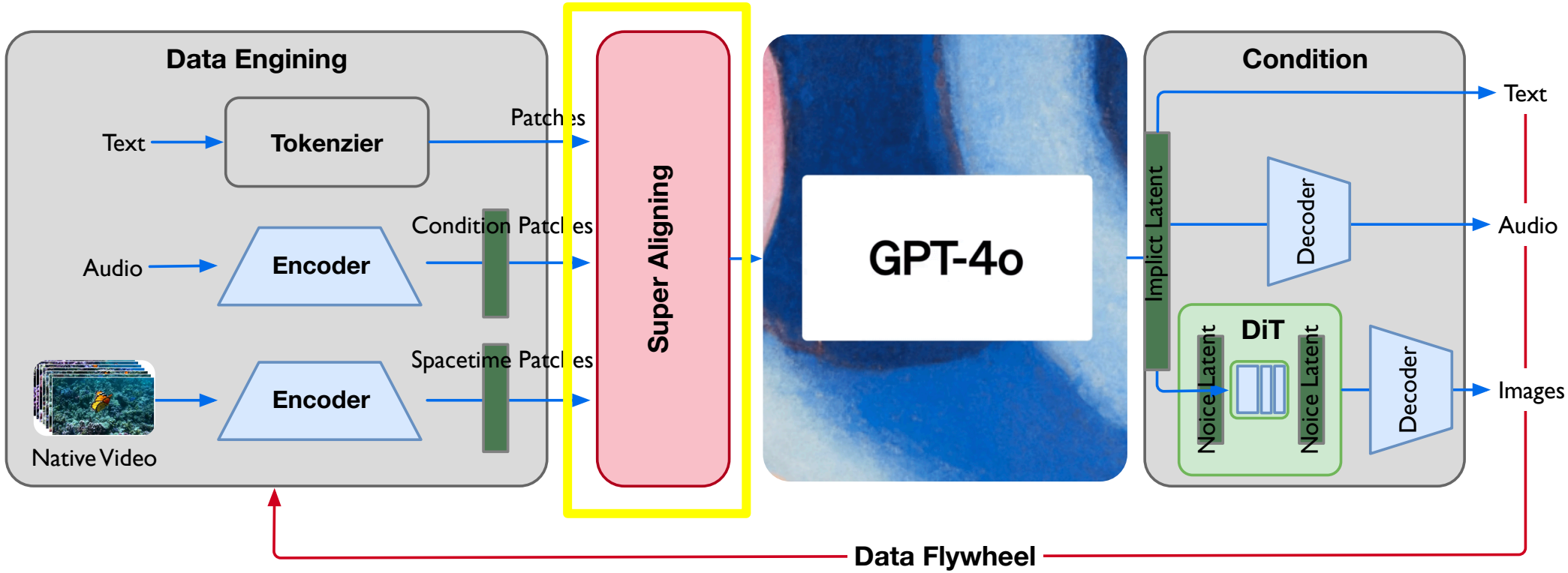
- Language tokenization：These 20 languages were chosen as representative of the new tokenizer‘s co mpression across different language families.

Gujarati 4.4x fewer tokens (from 145 to 33)	હેલો, મારું નામ જીપીટી-4o છે. હું એક નવા પ્રકારનું ભાષા મોડલ છું. તમને મળીને સારું લાગ્યું!
Telugu 3.5x fewer tokens (from 159 to 45)	నమస్కారము, నా పేరు జీపీటీ-4o. నేను ఒక్క కొత్త రకమైన భాషా మోడల్ ని. మిమ్మల్ని కలిసినందుకు సంతోషం!
Tamil 3.3x fewer tokens (from 116 to 35)	வணக்கம், என் பெயர் ஜிபிடி-4o. நான் ஒரு புதிய வகை மொழி மாடல். உங்களை சந்தித்ததில் மகிழ்ச்சி!
Marathi 2.9x fewer tokens (from 96 to 33)	नमस्कार, माझे नाव जीपीटी-4o आहे! मी एक नवीन प्रकारची भाषा मॉडेल आहे! तुम्हाला भेटून आनंद झाला!
Hindi 2.9x fewer tokens (from 90 to 31)	नमस्ते, मेरा नाम जीपीटी-4o है। मैं एक नए प्रकार का भाषा मॉडल हूँ। आपसे मिलकर अच्छा लगा!
Urdu 2.5x fewer tokens (from 82 to 33)	ہیلو، میرا نام جی پی ٹی-4o ہے۔ میں ایک نئے قسم کا زبان ماڈل ہوں، آپ سے مل کر اچھا لگا!

Korean 1.7x fewer tokens (from 45 to 27)	안녕하세요, 제 이름은 GPT-4o입니다. 저는 새로운 유형의 언어 모델입니다, 만나서 반갑습니다!
Vietnamese 1.5x fewer tokens (from 46 to 30)	Xin chào, tên tôi là GPT-4o. Tôi là một loại mô hình ngôn ngữ mới, rất vui được gặp bạn!
Chinese 1.4x fewer tokens (from 34 to 24)	你好，我的名字是GPT-4o。我是一种新型的语言模型，很高兴见到你!
Japanese 1.4x fewer tokens (from 37 to 26)	こんにちは、私の名前はGPT-4oです。私は新しいタイプの言語モデルです。初めまして！
Turkish 1.3x fewer tokens (from 39 to 30)	Merhaba, benim adım GPT-4o. Ben yeni bir dil modeli türüyüm, tanıştığımıza memnun oldum!
Italian 1.2x fewer tokens (from 34 to 28)	Ciao, mi chiamo GPT-4o. Sono un nuovo tipo di modello linguistico, piacere di conoscerti!

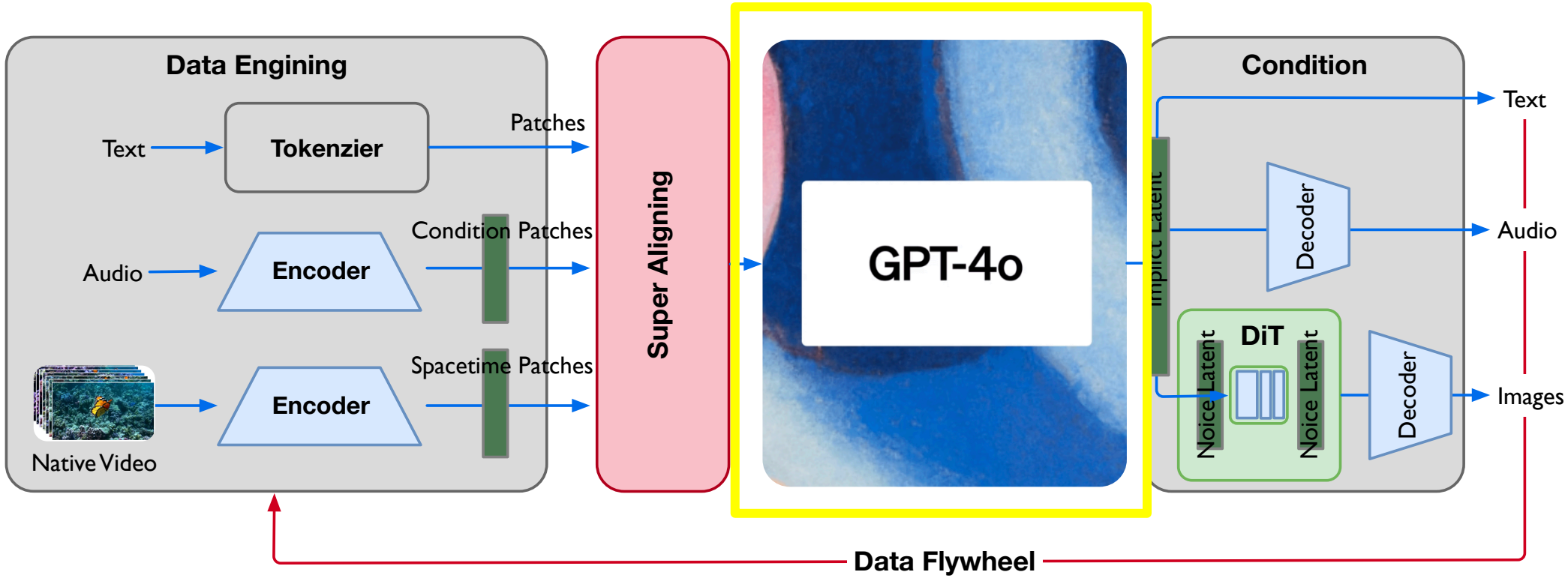
GPT-4o 原理解读：Super Aligning

- 端到端 E2E 的 MLM 大模型，对齐不同模态的输入，统一作为 Transformer 结构的长序列输入；
- 让能力弱的大模型监督能力强的大模型（ LLM supervise MLM ）；
- https://openai.com/index/introducing-superalignment/?utm_source=tldrai



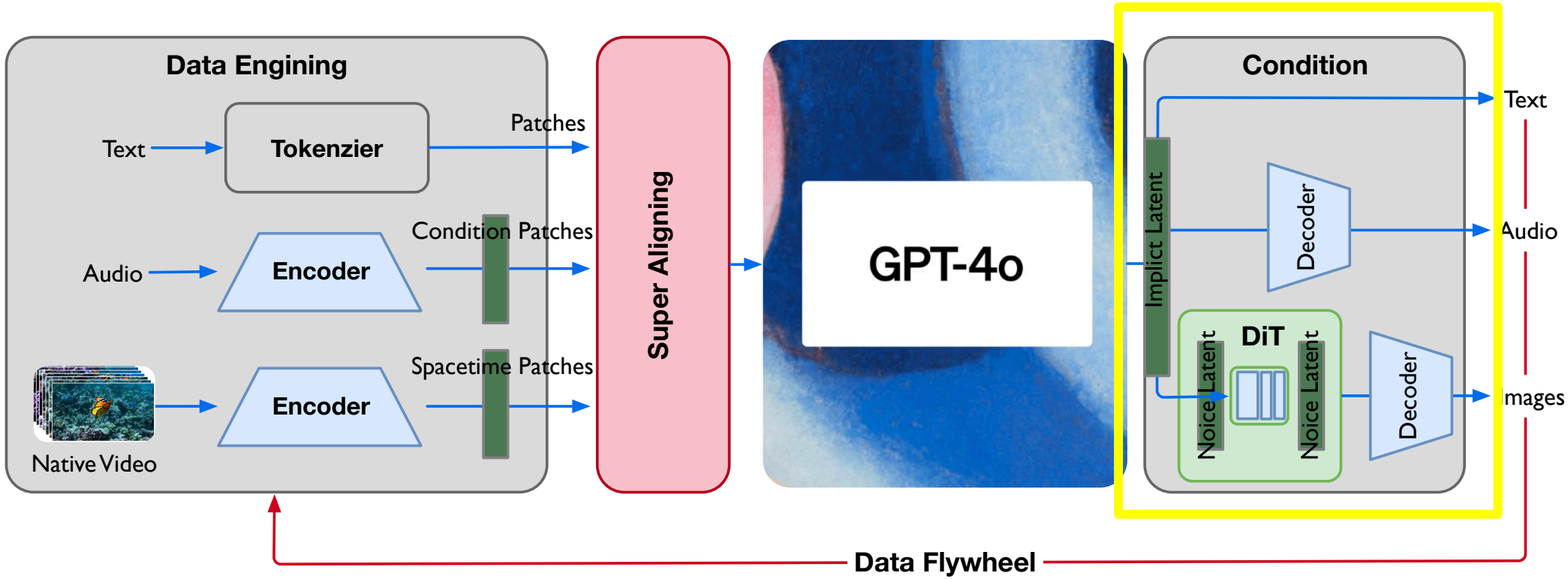
GPT-4o 原理解读：Transformer Decoder

- 纯 Transformer Decoder 架构，更加方便训练进行千卡、万卡规模的并行；
- 推理使用大融合算子（Flash Attention）进行极致加速；
- 符合 OpenAI 一贯 Everything Scaling Law 的方式；



GPT-4o 原理解读：Outputs

- 输出可配置、可选择 text/audio/images，因此是 Conducting 的case，统一 Transformers Tokens 输入可实现；
- Images 依然借鉴 SORA 使用 DiT 作为生成，但此处生成的为 Images not Videos；
- Audio 与 Text 应该会有对齐，保持同声传译；



GPT-4o 原理解读：Data flywheel

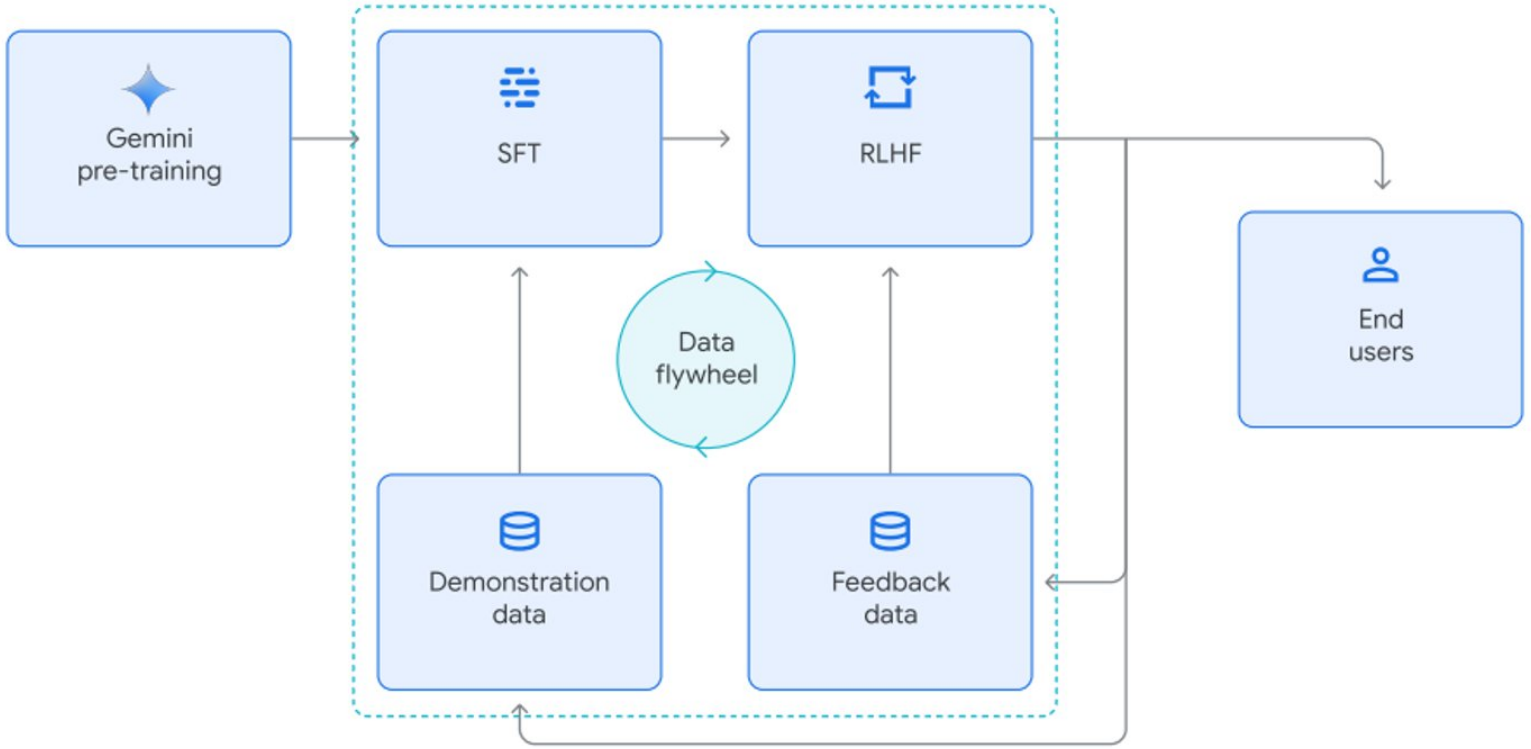


Figure 7 | **Modeling overview.** Post-training utilizes an optimized data flywheel in order to acquire human-AI feedback and continually improve on key areas. The data mixtures for supervised fine-tuning, reward modeling, and reinforcement learning serve as the foundation for our models.

技术总结

- **多模态数据工程：**
  1. LLM tokens 减少，让大模型的输入序列 Tokens 结合多模态统一为 Signal 长序列；
  2. 词表增大 Token 减少， 分词从 100K 到 200K，LLM 编码率进一步增强；
  3. Video 借鉴 SORA 对 spacetime patch 对时序极高编码率；
- **模型训练：**
  1. 弱监督/自监督为主，否则多模态对齐进行统一模式训练非常难；
- **模型结构与训练：**
  1. 通过 Super Aligning 对 Text、Audio、Video 三种模态进行对齐；
  2. 仍然以 LLM（GPT4）能力为主，加入多模态维度 Tokens 形成一个大模型；

4. 看大模型趋势

大模型发展趋势与思考

- 对 LLM 大模型的趋势思考；
- 对 MLM 大模型的趋势思考；
- 对百模大战厂商的思考（反思）；
- 对计算产业（AMD、NVIDIA、HUAWEI）的思考；

对大模型本身的思考

- LLM 大模型已经进入瓶颈期，加持 MOE 模型规模更大、超大词表、长序列：

1. DeepSeek、Grok 等对 MOE 的创新让模型规模更大，效果更好；
2. 如 KIMI 等追求长序列提升召回，但同时提供非长序列的 SFT/微调 系列模型；
3. OpenAI GPT-4o 架构在 LLM 微调整，但没有追求长序列而是：多模态、数据工程发展；
4. LLM 整体在 GPT-4 低 X% 精度波动，无突破性进展，进入应用下半场；

对大模型本身的思考

- MLM 进入赛道，24 年可称为多模态大模型元年，国外巨头在布局：

1. 国内多模态极少部分厂商在试水训练，GOOGLE 和 OpenAI 方向明确后下半年国内会涌现 MLM；
2. 大厂中高层容易被 PPT 忽悠，起大早赶晚集，手握万卡无 scaling law；
3. 上半年赛道压轴 SORA 而非 MLM，大模型赛道从 LLM 开始分裂；
4. MLM 围绕 LLM 的成果进一步升级迭代，而不是从新开始；

对百模大战产商思考

- 新赛道可以抄袭，但认知迭代速度不够快，很容易就被嘲讽：

1. Google 积极拥抱 AI，对自家搜索产品在逐步变革；百度用搜索防御者姿态参与 AI 竞争；
2. E2E 大模型 & E2E 大模型应用越来越多（Tesla FSD12、GPT-4o、Google Evo），国内串联；
3. 随 GPT-4 没重大提升， 国内 LLM 开始追赶上来，围绕 LLM 的具身智能走向落地探索阶段；
4. 国内外赛道差异明显，鲜见国内有 OpenAI 的 RoadMap 格局；

对计算产业的思考

- MLM、LLM 等大模型越来越多，服务器算力持续消耗，推理算力提前布局：

1. 24 年仍然仍然是重服务器算力，持续消耗大量算力，厂商对算力的渴求仍然不够；
2. 推理芯片继续加强布局（1/2 年代际），等待 MLM、LLM 等应用成熟落地，but LLM 芯片 XXX？
3. MLM 对特殊解码器（IP）有新需求，数据工程不满足于 ARM CPU 的小核小 DIE 方式；
4. MLM 推理高吞吐下低时延 vs LLM 推理低时延高吞吐，HBM or LPDDR？

Search ⌘ K

GET STARTED

Introduction

Quickstart

Models

Overview

Model updates

GPT-4o

GPT-4 Turbo and GPT-4

GPT-3.5 Turbo

DALL·E

TTS

Whisper

Embeddings

Moderation

GPT Base

How we use your data

Models overview

The OpenAI API is powered by a diverse set of models with different capabilities and price points. You can also make customizations to our models for your specific use case with fine-tuning.

MODEL	DESCRIPTION
GPT-4o	The fastest and most affordable flagship model
GPT-4 Turbo and GPT-4	The previous set of high-intelligence models
GPT-3.5 Turbo	A fast, inexpensive model for simple tasks
DALL·E	A model that can generate and edit images given a natural language prompt
TTS	A set of models that can convert text into natural sounding spoken audio
Whisper	A model that can convert audio into text
Embeddings	A set of models that can convert text into a numerical form
Moderation	A fine-tuned model that can detect whether text may be sensitive or unsafe
GPT base	A set of models without instruction following that can understand as well as generate natural language or code
Deprecated	A full list of models that have been deprecated along with the suggested replacement

参考文献 Reference

1. Chan, W., Park, D., Lee, C., Zhang, Y., Le, Q., and Norouzi, M. SpeechStew: Simply mix all available speech recogni- tion data to train one large neural network. arXiv preprint arXiv:2104.02133, 2021(opens in a new window).
2. 2Galvez, D., Diamos, G., Torres, J. M. C., Achorn, K., Gopi, A., Kanter, D., Lam, M., Mazumder, M., and Reddi, V. J. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. arXiv preprint arXiv:2111.09344, 2021(opens in a new window).
3. 3Chen, G., Chai, S., Wang, G., Du, J., Zhang, W.-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. arXiv preprint arXiv:2106.06909, 2021(opens in a new window).
4. 4Baevski, A., Zhou, H., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. arXiv preprint arXiv:2006.11477, 2020(opens in a new window).
5. 5Baevski, A., Hsu, W.N., Conneau, A., and Auli, M. Unsu pervised speech recognition. Advances in Neural Information Processing Systems, 34:27826–27839, 2021.
6. 6Zhang, Y., Park, D. S., Han, W., Qin, J., Gulati, A., Shor, J., Jansen, A., Xu, Y., Huang, Y., Wang, S., et al. BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. arXiv preprint arXiv:2109.13226, 2021(opens in a new window).

Thank you

把AI系统带入每个开发者、每个家庭、每个组织，构建万物互联的智能世界

Bring AI System to every person, home and organization for a fully connected, intelligent world.

ZOMI

Course chenzomi12.github.io

GitHub github.com/chenzomi12/DeepLearningSystem