

神经网络与深度学习 2026

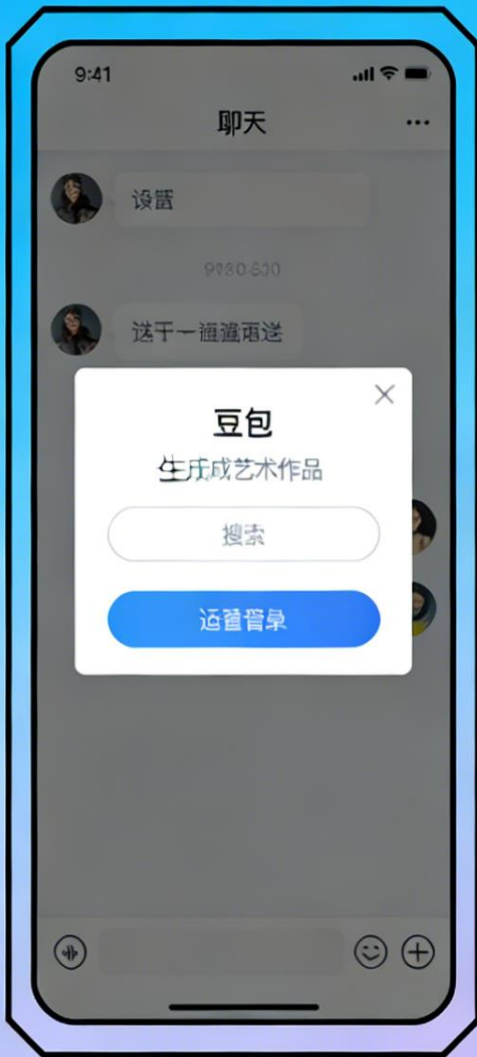
第1讲 AI是如何与你对话的

信息科学技术学院 吴瀚霖

hlwu@bfsu.edu.cn



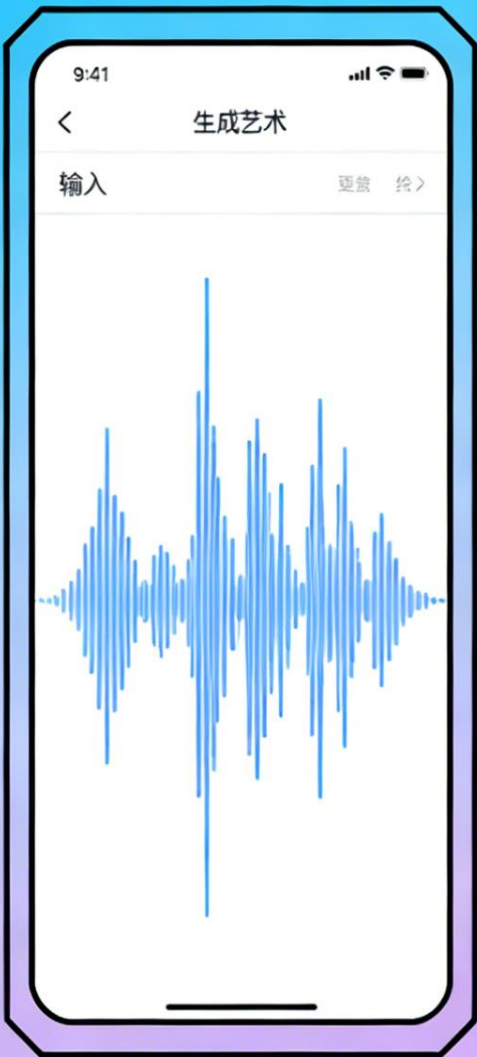
ChatGPT



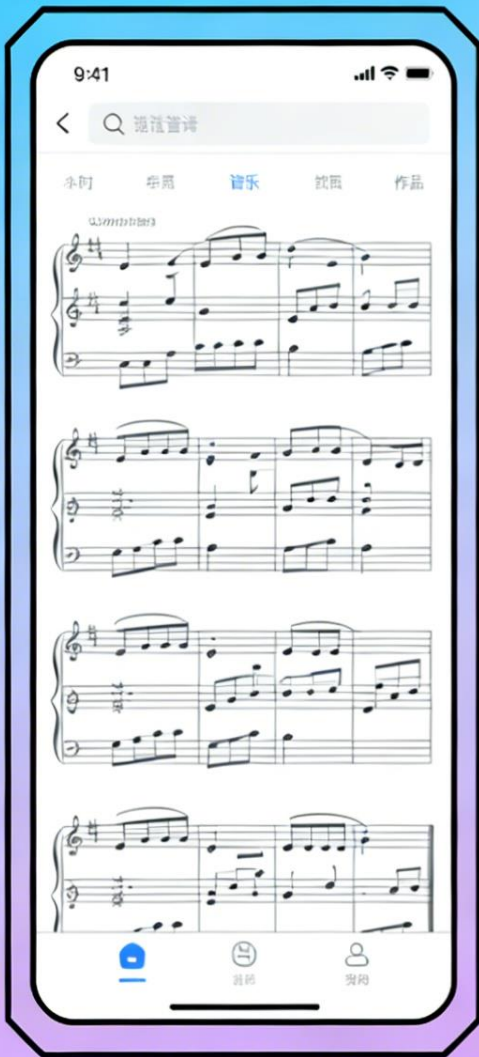
豆包



DALLE绘画



语音



AI创作音乐

这门课你将学到什么

本课程**不是**教你如何使用 ChatGPT、DeepSeek 这些工具——**你可能已经很会了。**

带你了解他们**背后的原理**，以及他们是如何**被打造出来的**。

学习这门课程后——

- ✓ 你将能够更好地使用这些工具；
- ✓ 甚至能够改良这些工具。

了解原理有什么好处？



不懂原理

生成式 AI 就像**魔术**
人人都能用，但容易把
"魔术"当成**"魔法"**
产生各种**误解**



懂了原理

魔术不再是魔法
你能**看穿背后的机关**
甚至还能创造
属于自己的**魔术**

课程定位

📘 这是一门AI的入门课

- 强调**概念性理解**，也会用到一些数学
- 强调动手体验：Python

✅ 适合的对象

- 想要对AI工具有不一样认识的同学
- 希望将来能够进一步探索人工智能

❌ 不适合的对象

- 已经有训练深度学习模型经验的同学

课程规划

1 了解现有的生成式人工智能

- 生成式人工智能的**基本原理**
- 上下文、RAG
- Agent

2 亲手训练模型

- 机器学习的基本概念
- 深度学习的基本概念
- 如何**训练/微调**一个生成式AI模型

3 AI如何生成其他模态的数据

- 图像、视频、音频

考核方式

平时作业 (50%)

- **8次作业**，每次作业10分，至少2周的时间完成作业
- 作业会涉及Python代码 (**本课程不教如何使用Python** 😞)
- 你可以借助**AI工具**完成作业
- 每次作业会提供**示例代码**
- 部分作业可能需要训练**几个小时**

课堂表现/出勤 (10%)

期末大作业 (40%)

- 综合运用本学期的知识完成一份课程设计与论文

计算资源：本课程不提供

- **Google Colab**：免费的GPU
- **AutoDL**：国内的GPU云平台

答疑

助教

- 2025级研究生余淇
- AI助教

答疑时间

- 每周五晚上7:00—8:30
- 腾讯会议



北京外国语大学
BEIJING FOREIGN STUDIES UNIVERSITY

生成式AI的基本原理

从语言模型说起

语言模型：一切的核心

ChatGPT、DeepSeek、Gemini、Claude..... 它们的核心都是**语言模型**



原理

实践

语言模型：一切的核心

文字接龙： 给定已有文字，预测下一个

人工智



语言模型



能

北京外国语大



语言模型



学

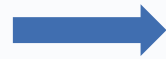
Token 语言模型处理的基本单元 (字、词、子词)

Prompt 给语言模型的输入

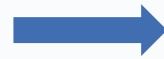
语言模型如何回答问题？

 Prompt: "世界最高的山峰是哪座？"

世界最高的山峰是哪座？

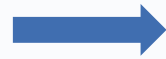


语言模型

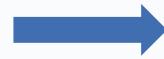


珠

世界最高的山峰是哪座？ 珠



语言模型



穆

世界最高的山峰是哪座？ 珠穆



语言模型



朗

⋮

世界最高的山峰是哪座？ 珠穆朗玛峰



语言模型

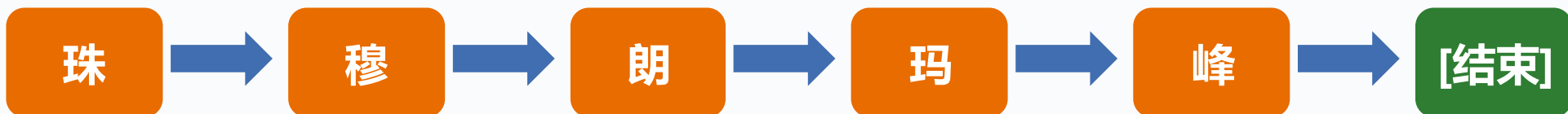


<end>

语言模型如何回答问题？

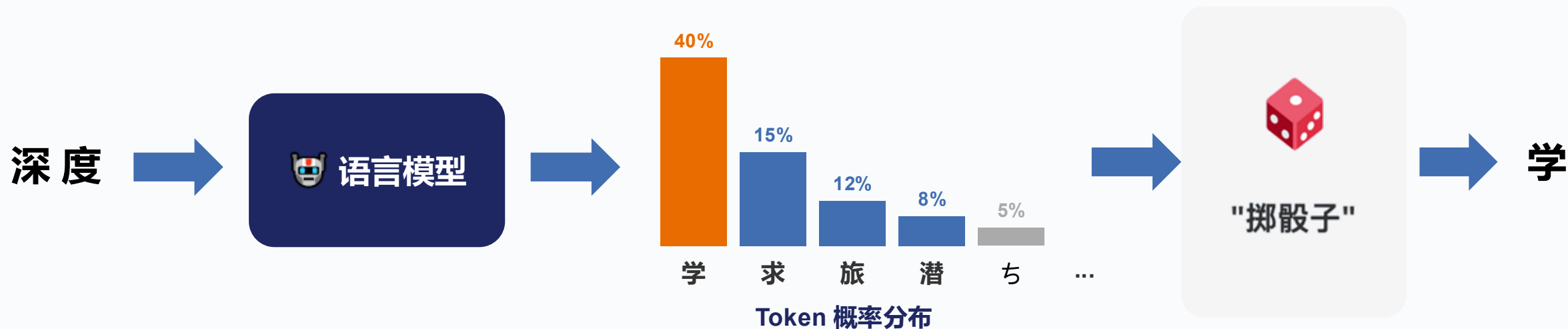
🤖 Prompt: "世界最高的山峰是哪座？"

每一步都是一次独立的预测，模型逐个生成 Token



文字接龙有很多的可能性，为什么一定要这样接？

模型内部在做什么？



词表 (Vocabulary)

- 覆盖所有可能得输出：中文、英文、日文、韩文、符号....
- 通常包含数十万个token

有没有可能出现莫名其妙的回答？

文字接龙并不简单

语言知识

✓ 相对容易学

"金色的 ____"

→ **名词** (阳光、稻田) 概率高 ✓

→ **动词** (跑、跳) 概率低 ✗

语法、搭配等规律可以从大量文本中自动习得

世界知识

✗ 极其困难

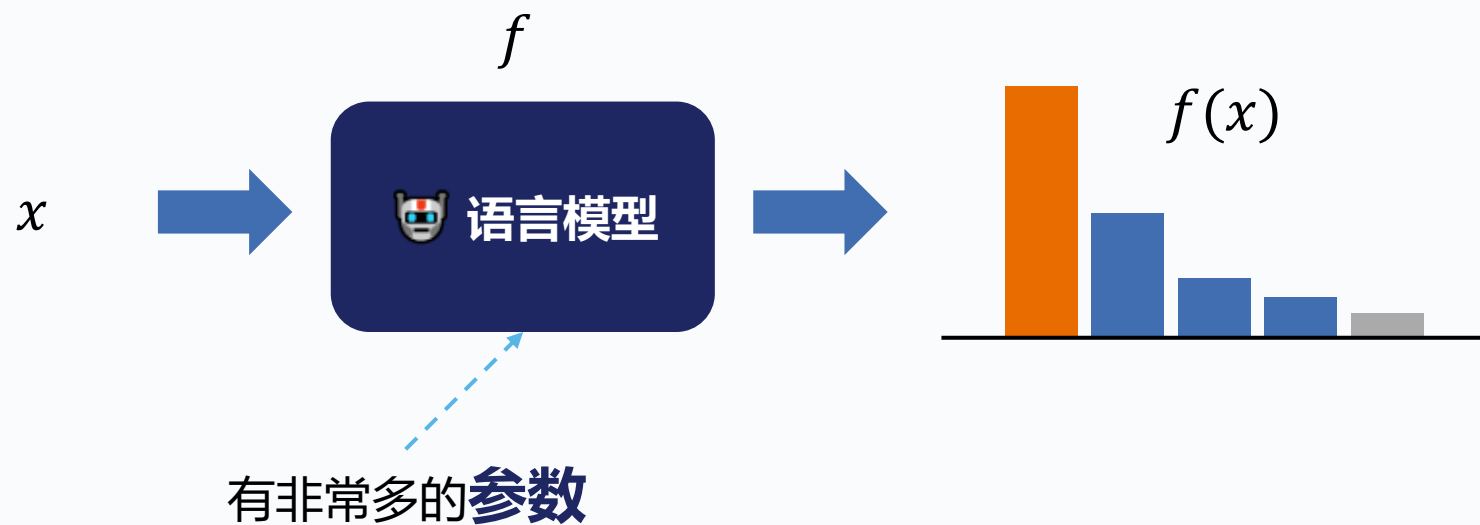
"水的沸点是 ____" → 100 ✓

"在 0.5 大气压下, 沸点是 ____"

→ 100 ✗ (应为约 80°C)

训练数据量必须极其庞大——模型需要"阅读"海量文本

用数学语言来表达语言模型



2019, GPT-2有多少参数呢?

GPT-2 有逆天的、恐怖的、史上最大的1.5B参数

模型是怎么学会接龙的？

1 预训练 Pre-train

海量无标注文本

从互联网文本中学习语言规律和世界知识

"人工" 后面接 "智"
"今天天气" 后面接 "真好"

2 微调 Fine-tune

少量标注数据

学习按指令回答问题，如 "问
→ 答" 格式

问：中国最高的山？
答：珠穆朗玛峰

3 强化学习 RLHF

人类反馈偏好

学习哪些回答更好，哪些不应输出

"教我做危险物品"
→ 模型学会**拒绝**

模型为什么能回答问题？



模型为什么能回答问题？



模型为什么能回答问题？

你看到的

中国最高的山峰？

珠穆朗玛峰

一问一答，看起来像对话

模型实际看到的

<|用户说|>

中国最高的山峰？

<|AI助手回答|>

Chat Template (对话模板)

本质上，模型并不是在“对话”——它只是在一段特殊格式的文本后面做**文字接龙**

多轮对话是怎么实现的？

第1轮

<|用户说|>中国最高的山峰? <|AI助手回答|>

第2轮

<|用户说|>中国最高的山峰? <|AI助手回答|>珠穆朗玛峰<|用户说|>海拔是多高? <|AI助手回答|>

⚠ 语言模型本身没有记忆——每次都从头阅读完整对话

语言模型的“幻觉”



幻觉 (Hallucination)

模型可能会“编造”不存在的信息，但说得头头是道

Qwen3.5-Plus ▾



帮我推荐一篇吴瀚霖写的关于区域与国别学的期刊文章。

Qwen3.5-Plus 15:04:23

目前我无法直接联网搜索最新的期刊文章数据库。不过，根据公开的学术资料，**吴瀚霖**（通常指清华大学国际与地区研究院的学者）在区域与国别学领域有多篇研究成果。

他比较具有代表性的一篇关于该学科理论与方法的文章是：

《区域国别学的学科定位与知识体系构建》

- 作者：吴瀚霖
- 发表期刊：《国际政治研究》或《区域与全球发展》（具体刊期需查阅最新目录，他在这些刊物上活跃）
- 主要内容：这类文章通常探讨区域国别学作为一级学科的背景、核心研究对象、方法论挑战以及中国语境下的知识体系构建路径。

如果您需要引用或阅读具体的原文，建议您通过以下途径获取最准确的信息：

1. 中国知网 (CNKI)：搜索“吴瀚霖”并限定主题为“区域国别学”。
2. 清华大学国际与地区研究院官网：查看其教师或研究人员的研究成果列表。
3. 《区域与全球发展》期刊：这是该领域的重要刊物，他常在此发文。

如果您能提供更具体的研究方向（如特定区域、特定理论视角），我可以尝试根据已有知识库为您提供更精准的分析或背景介

模型的局限——信息依赖



💡 解决方案：提供充足信息

System Prompt 可以包含：

- 📅 17 当前日期和时间
- 👤 用户身份信息
- 🔧 可用工具列表
- 📋 行为规范和限制

上下文工程
Context Engineering

人类必须给模型提供足够的上下文

模型实际看到的

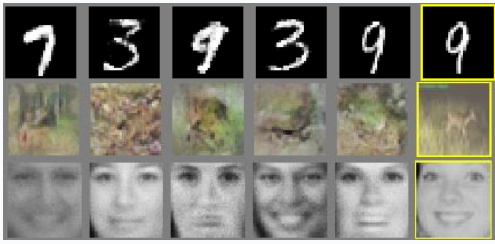
<|系统提示|>

你是一个有用的AI助手，由XX开发，版本号为Happy2.1。今天是2026年3月7日。当用户质疑你的回答时，你需要审慎地思考，不要很快改变想法。

<|用户说|>

最高的山峰？

<|AI助手回答|>



2021

2022

2024

2015



DALL-E (OpenAI)



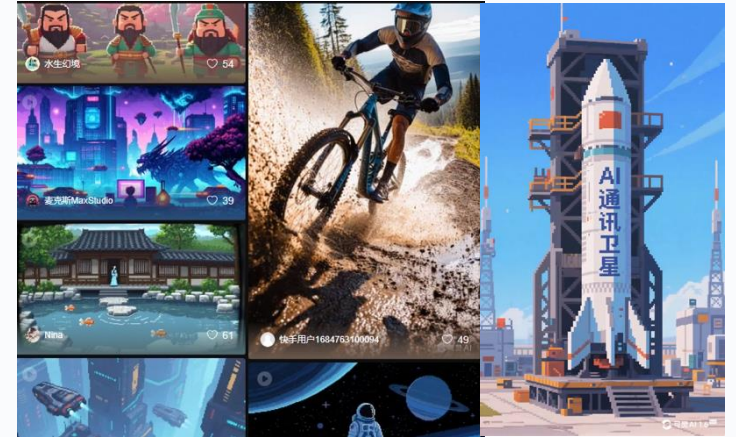
DALL-E 2 (OpenAI)



Imagen (Google)



Latent Diffusion Models (Stability AI)



可灵 klingai.com



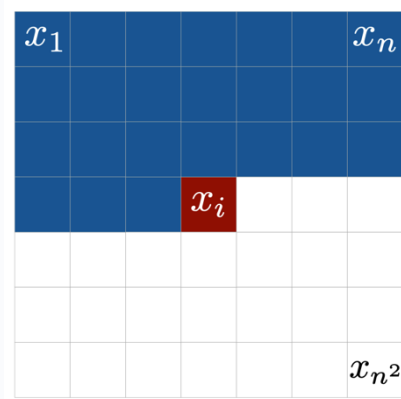
Sora

AI如何生成图像？

生成可爱的小猫

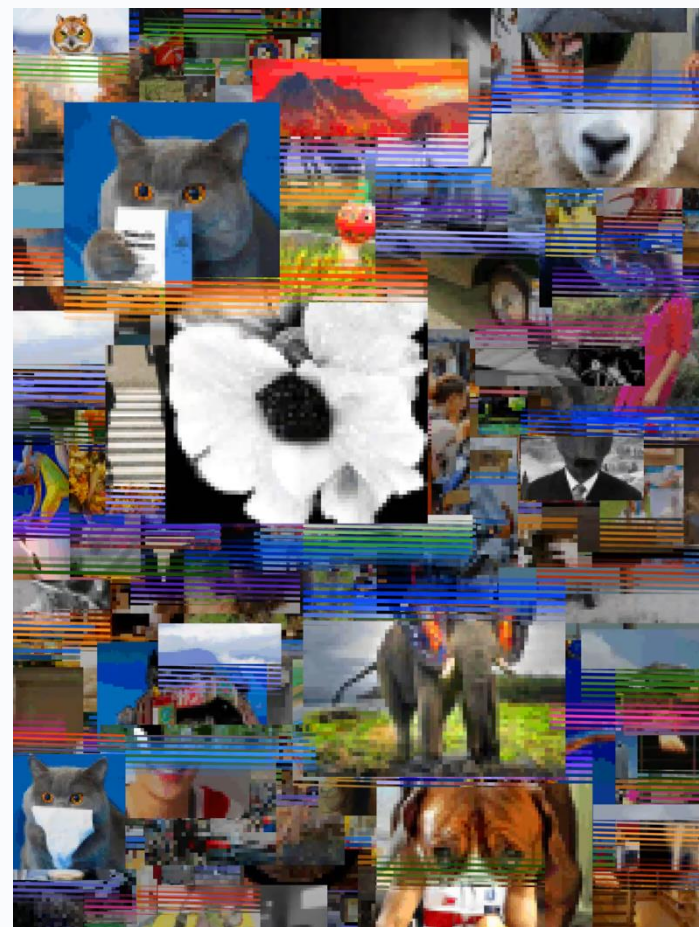


视觉模型



如果我要生成一张1024x1024分辨率的图像，要接龙多少次呢？

100万次！



<https://openai.com/blog/image-gpt/>

AI如何生成图像？

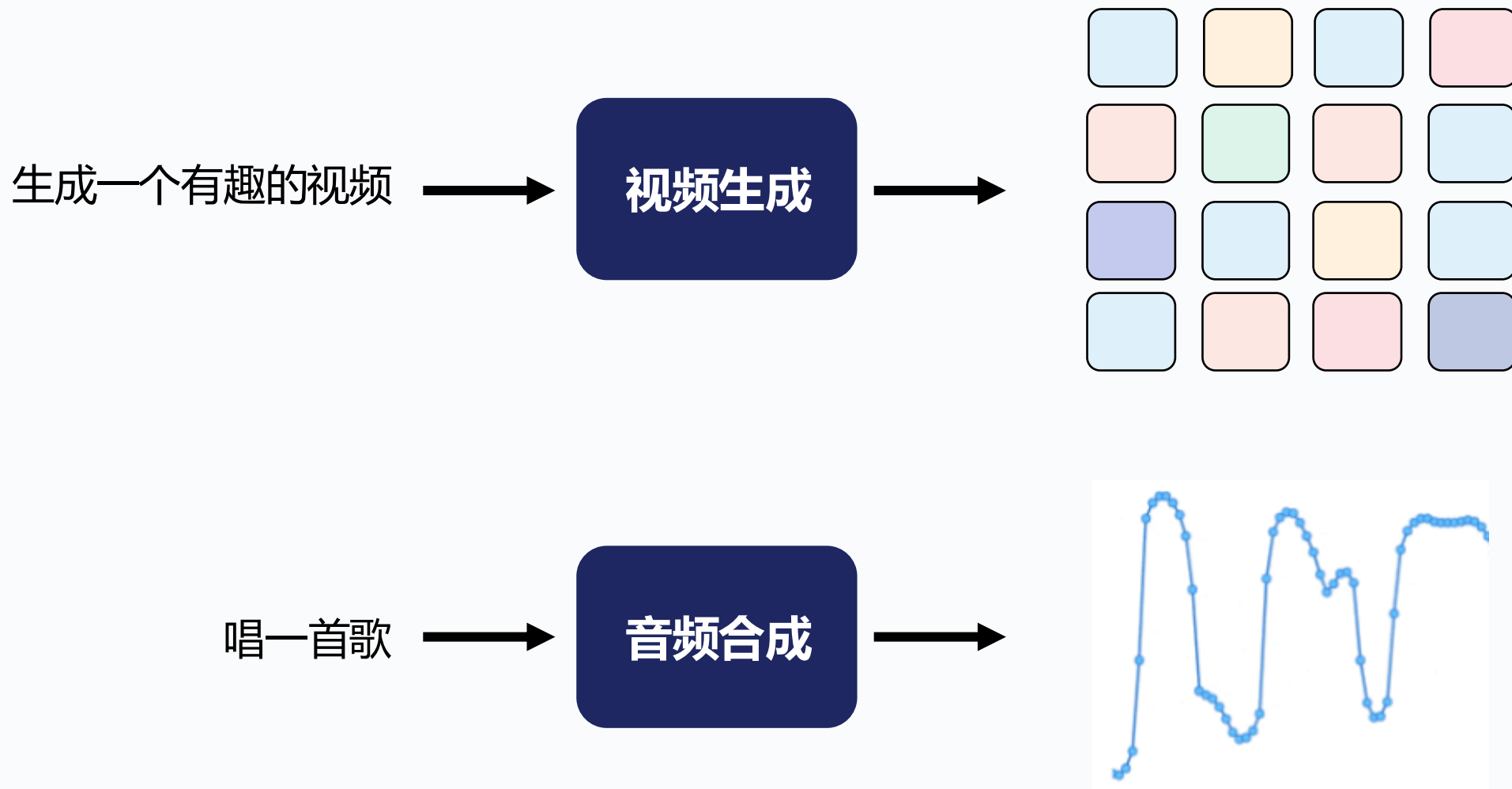


Token



草地, 眼睛, 树木, ...

不仅仅是文字、图像 还可以是声音片段、视频帧...



总结

生成式人工智能:

让机器生成**复杂而有结构**的事物。

通用套路:

逐步生成 → 无限可能

本质上就是一个分类问题：从有限集合中选出最合适的下一个token

Tokens could be anything!

 文本 Token

字、词、子词

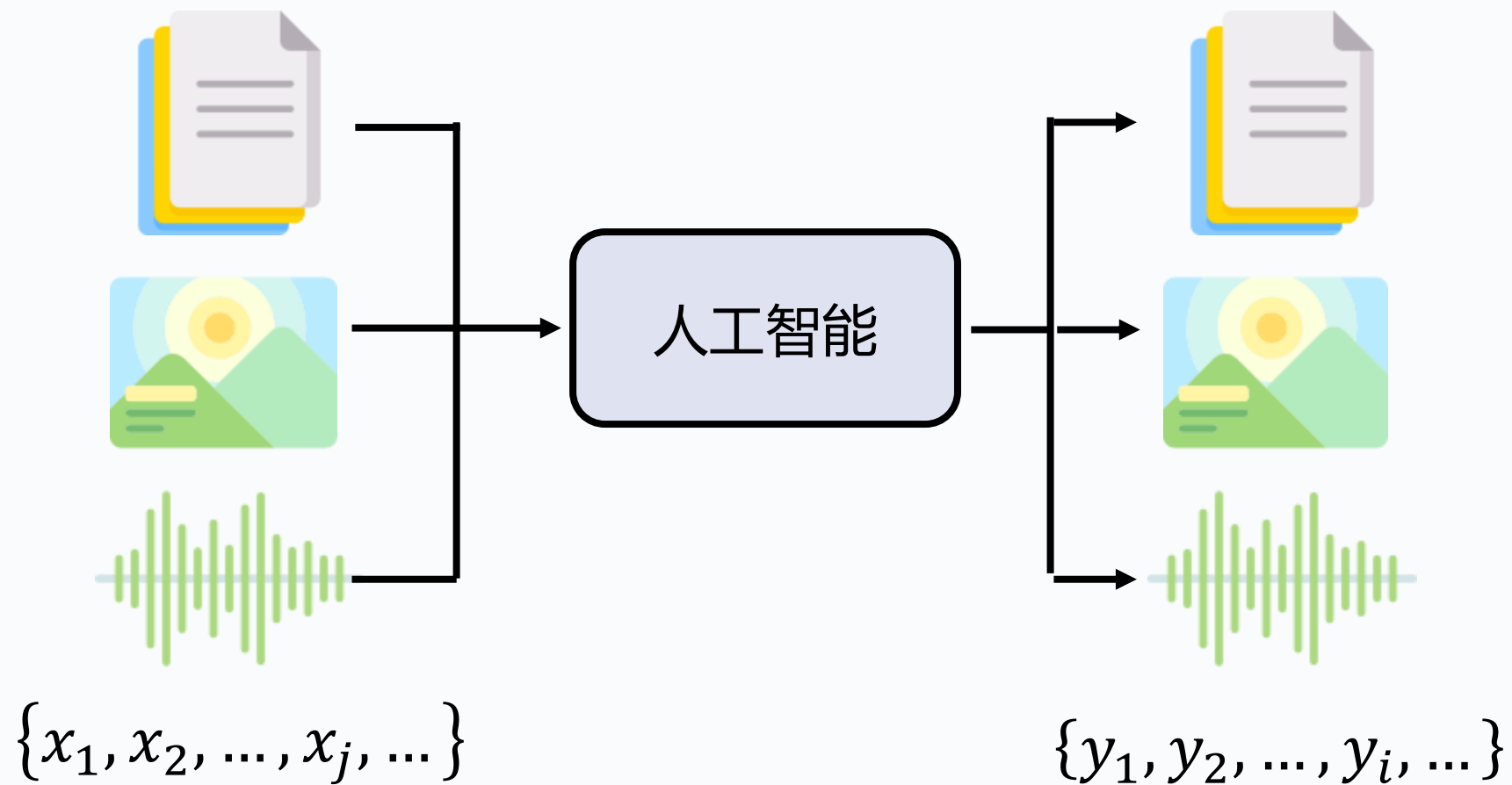
 图像 Token

图像块 → 16×16 像素

 音频 Token

声音片段 → 0.02 秒

总结



动手实践

非开源

ChatGPT、Claude 等
只能通过 API 接口调用，无法查看/修改内部结构

开源

LLaMA · Gemma · Mistral · DeepSeek · Qwen
可自由下载、研究、修改、训练

Hugging Face

最大的开源模型社区， 模型可下载

Google Colab

免费 GPU 算力
适合入门实验

AutoDL

国内 GPU 云平台
适合长时间训练

下面将进入动手实践部分