大模型系列之智能体

大模型与智能体

ZOMI

关于大模型系列

- 内容背景
  - LLM + AI Agent：大模型遇到智能体
- 具体内容
  1. AI Agent 组成介绍：LLM + 记忆 + 规划 + 工具
  2. AI Agent 规划手段： Task Decomposition 与 Self Reflection
  3. AI Agent 热门应用：交互式 Agent、自动化 Agent 与多模态 Agent
  4. AI Agent 问题与挑战： Agent 的问题、Agent 的局限性

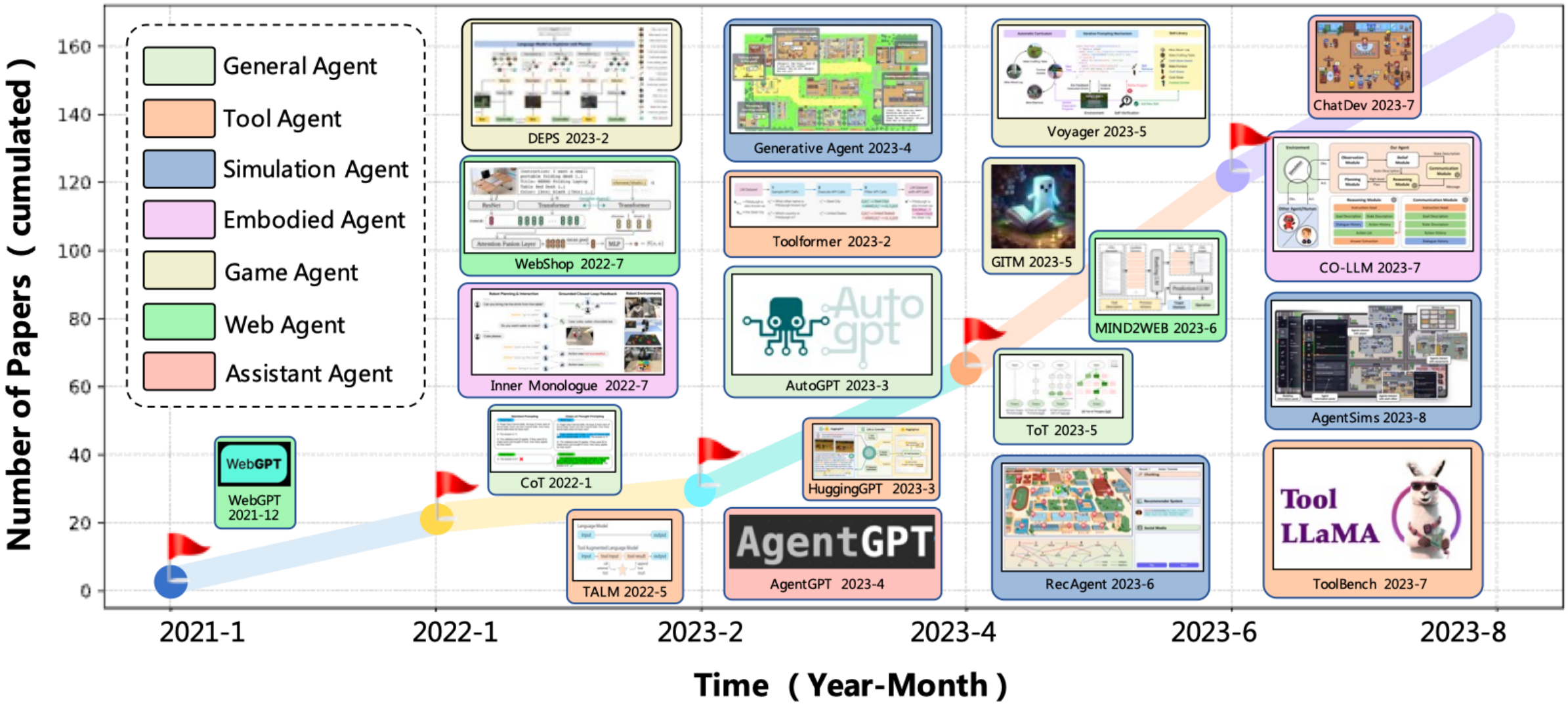
欣赏视频吧

- 《头号玩家》
- 《她》

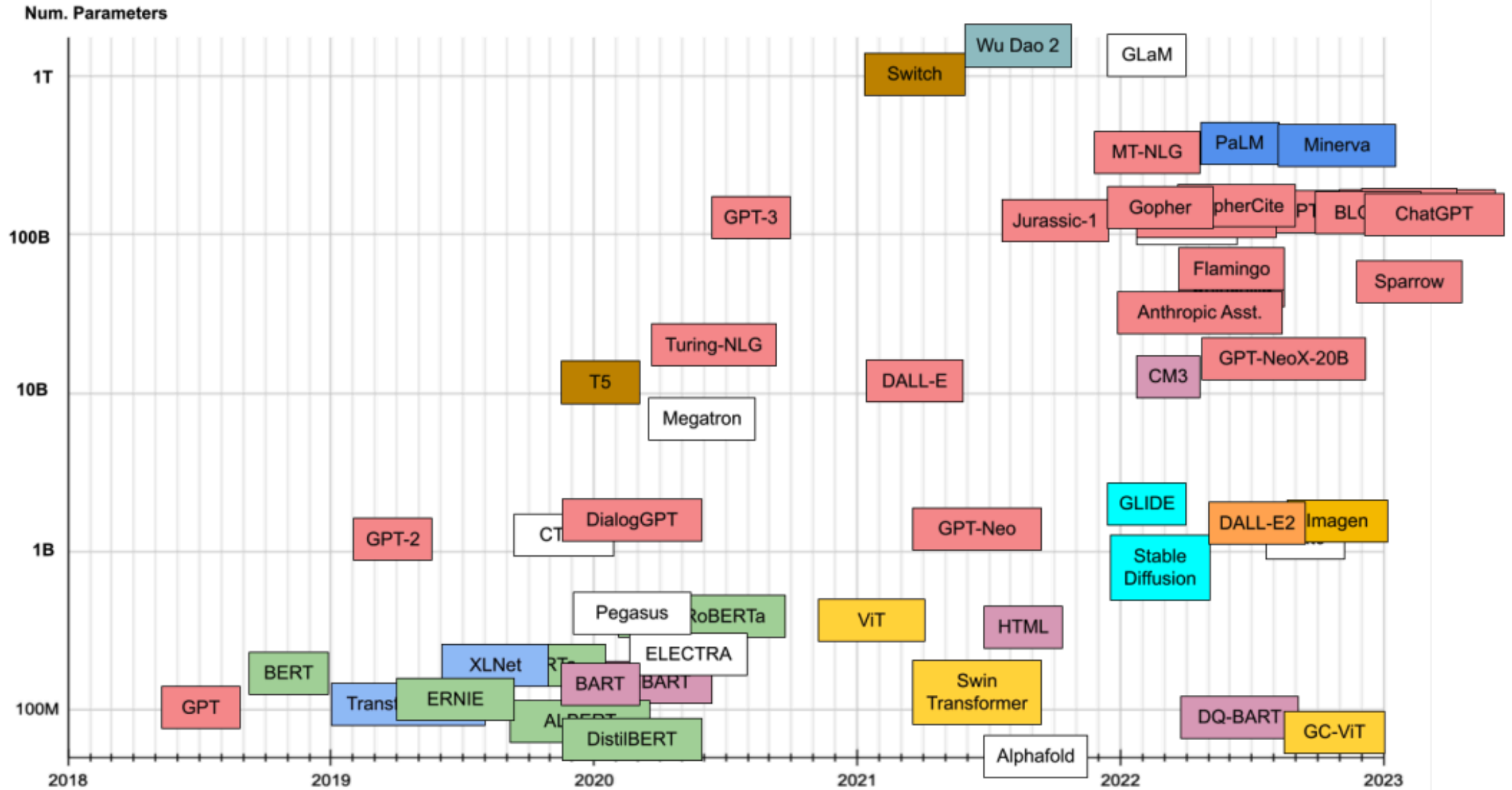
AI Agent 大爆发

2023 年各类智能体纷纷发布

- **3月21日，Camel 发布**：通过角色扮演自主AI智能体来解决给定的任务
- **3月30日，AutoGPT 发布**：根据任务设定自动地分阶段推进执行，并最终给出结果
- **4月3日，BabyAGI 发布**：人工智能驱动的任务管理系统
- **4月7日， Smallville 西部小镇发布**：斯坦福 25 个 AI 智能体相互交流
- **9月5日，ChatDev 发布**：清华创建一个由多智能体协作运营的虚拟软件公司
- **5月27日， Voyager发布**：英伟达 AI 智能体自主写代码，游戏中全场景终身学习
- **8月9日，OlaGPT 发布**：首个模拟人类认知的思维框架
- **10月29日，MetaGPT 发布**： 实现多智能体通信，智能体也能轻松狼人杀


General Agent
Tool Agent
Simulation Agent
Embodied Agent
Game Agent
Web Agent
Assistant Agent
Number of Papers (cumulated)
0
20
40
60
80
100
120
140
160
2021-1
2022-1
2023-2
2023-4
2023-6
2023-8
Time (Year-Month)
WebGPT
WebGPT 2021-12
DEPS 2023-2
WebShop 2022-7
Inner Monologue 2022-7
CoT 2022-1
TALM 2022-5
Generative Agent 2023-4
Toolformer 2023-2
AutoGPT 2023-3
HuggingGPT 2023-3
AgentGPT
AgentGPT 2023-4
Voyager 2023-5
GITM 2023-5
ToT 2023-5
RecAgent 2023-6
MIND2WEB 2023-6
ChatDev 2023-7
CO-LLM 2023-7
AgentSims 2023-8
Tool LLaMA
ToolBench 2023-7

1. LLM & Agent

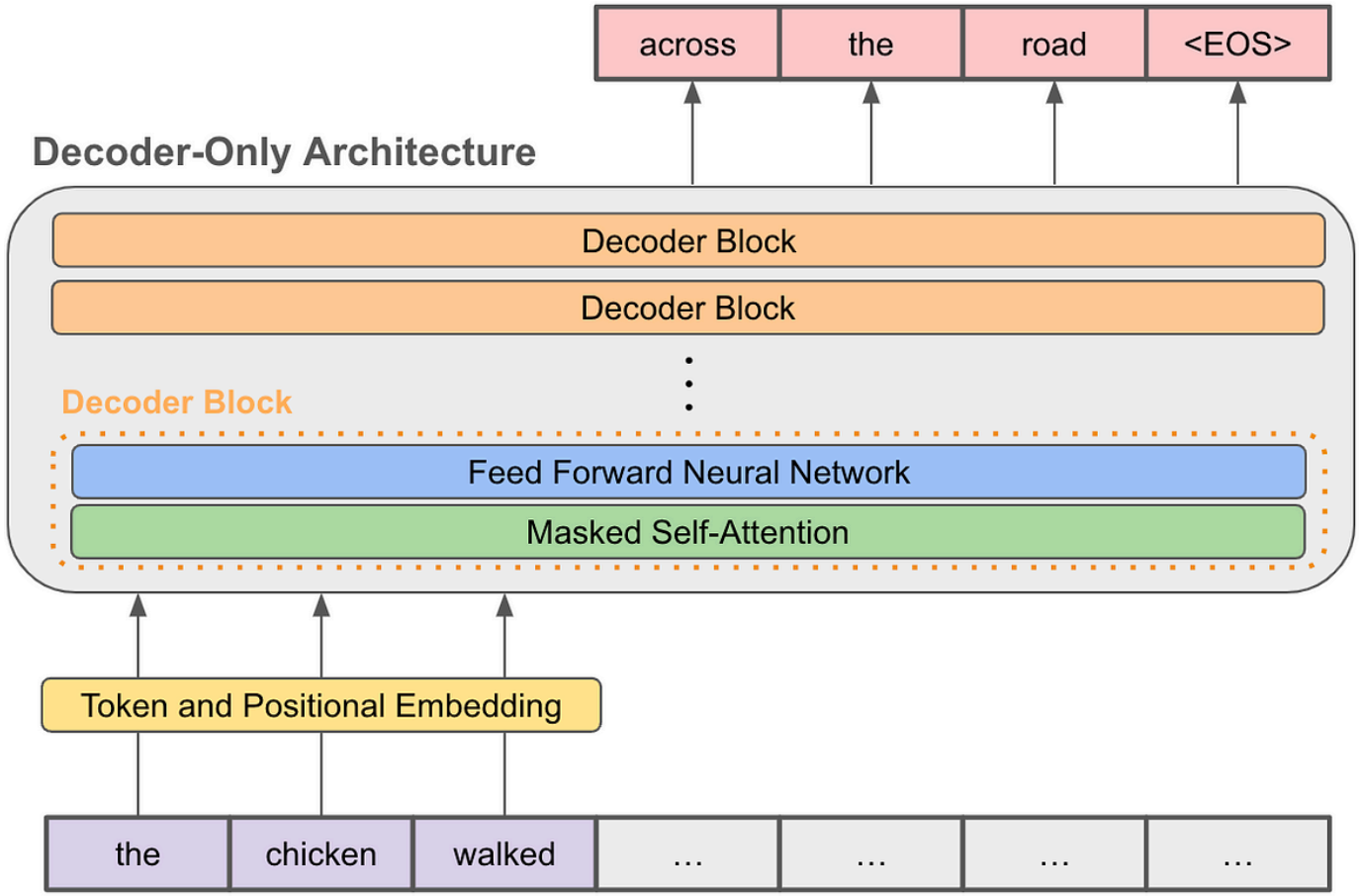

Num. Parameters
1T
100B
10B
1B
100M
2018
2019
2020
2021
2022
2023
Wu Dao 2
Switch
GLaM
MT-NLG
PaLM
Minerva
GPT-3
Jurassic-1
Gopher
ChatGPT
Flamingo
Sparrow
Anthropic Asst.
Turing-NLG
T5
DALL-E
CM3
GPT-NeoX-20B
Megatron
GLIDE
DALL-E2
Imagen
DialogGPT
GPT-2
GPT-Neo
Stable Diffusion
Pegasus
ViT
HTML
ELECTRA
XLNet
BERT
BART
Swin Transformer
GPT
ERNIE
DQ-BART
GC-ViT
DistilBERT
Alphafold

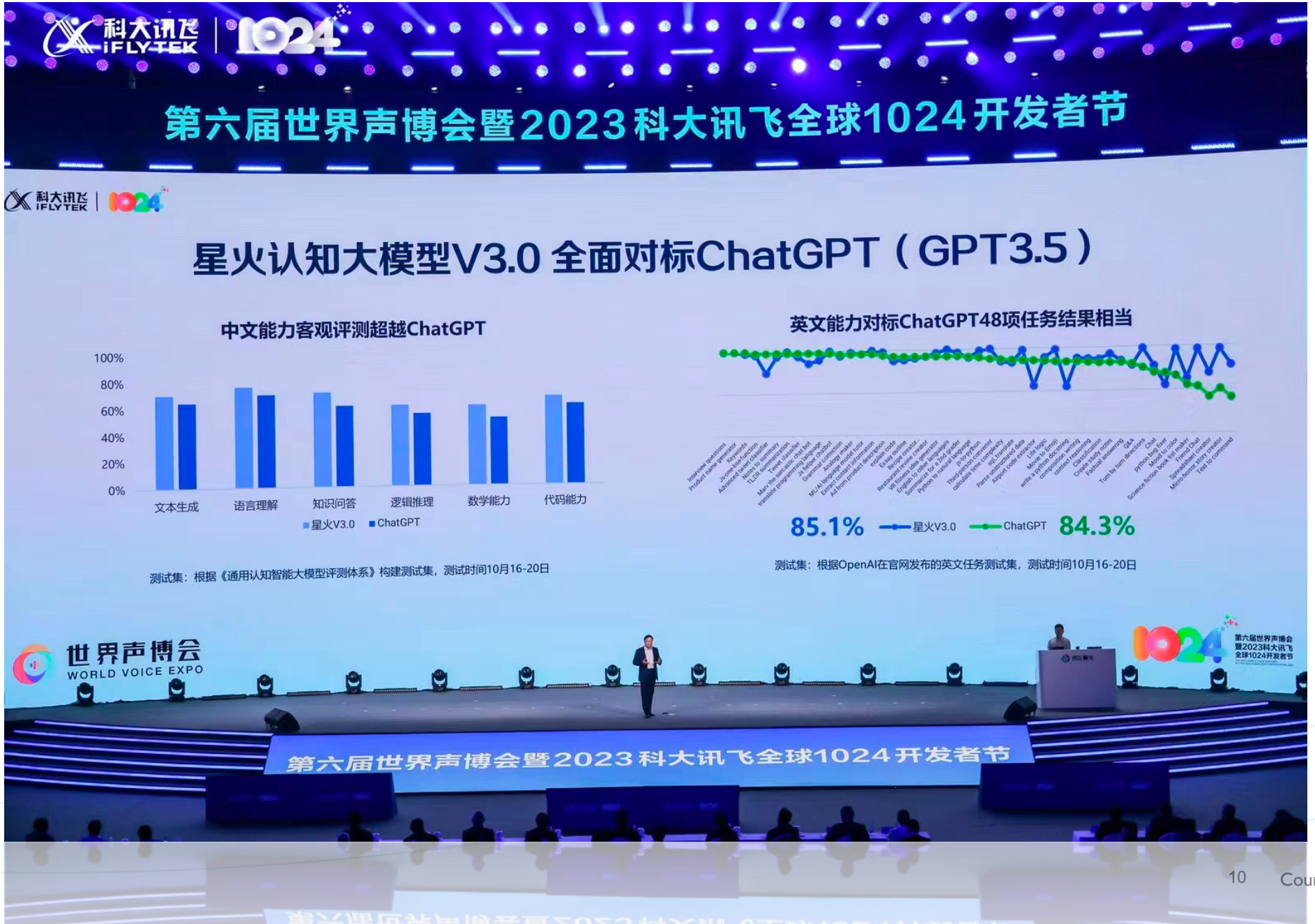
LLM to AGI 通用人工智能

- LLM 初期，对其能力边界还没有清晰认知，以为 LLM 是通向 AGI 的路径，即：LLM to AGI

LLM 回顾

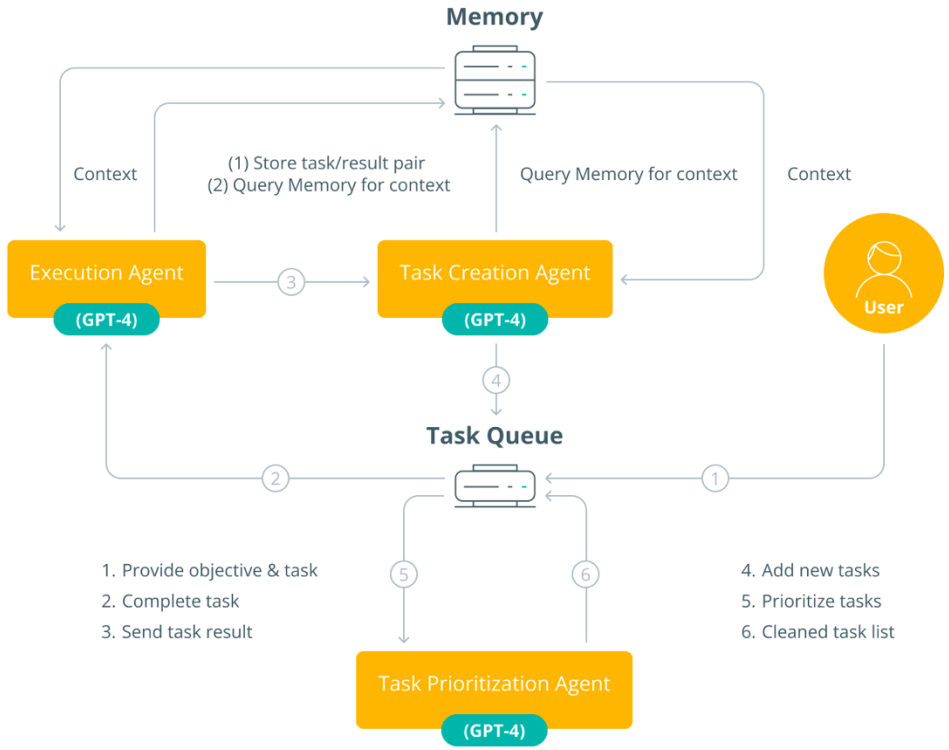
- 大语言模型（LLM）只能响应用户的查询指令，实现一些生成任务，比如写故事、生成代码等。




科大讯飞 iFLYTEK | 1024
第六届世界声博会暨2023科大讯飞全球1024开发者节
科大讯飞 iFLYTEK | 1024
星火认知大模型V3.0 全面对标ChatGPT（GPT3.5）
中文能力客观评测超越ChatGPT
100%
80%
60%
40%
20%
0%
文本生成
语言理解
知识问答
逻辑推理
数学能力
代码能力
星火V3.0
ChatGPT
测试集：根据《通用认知智能大模型评测体系》构建测试集，测试时间10月16-20日
英文能力对标ChatGPT48项任务结果相当
85.1%
星火V3.0
ChatGPT
84.3%
测试集：根据OpenAI在官网发布的英文任务测试集，测试时间10月16-20日
世界声博会
WORLD VOICE EXPO
第六届世界声博会暨2023科大讯飞全球1024开发者节

从 LLM 到 LAM

- AutoGPT 和 BabyAGI 等项目为代表的大动作模型（Large-Action Models / Large-Agent Models，LAM ）将 LLM 作为 Agent 的中心，将复杂任务进行分解，在每个子步骤实现自主决策和执行。



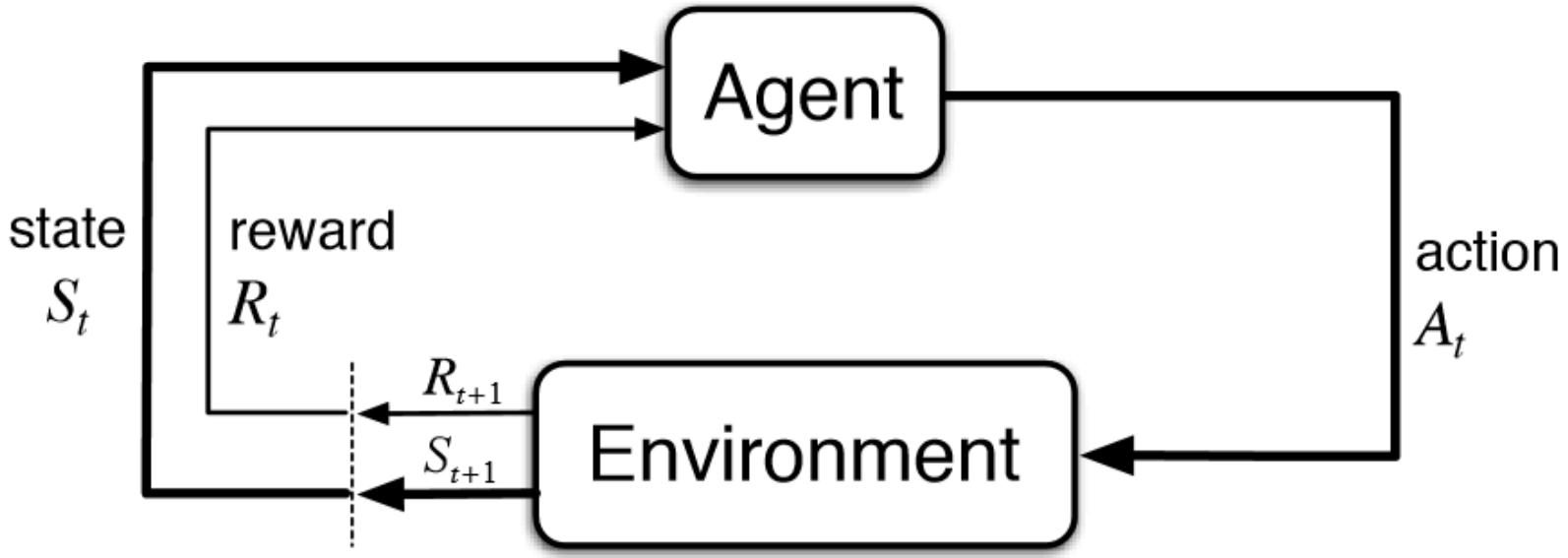
LLM to AI Agent to AGI

- Times later，LLM 既有问题并不能直接到达 AGI，但是借助一个/多个 Agent 可以有望实现通往 AGI 的道路：将最重要的“任务规划” 交由 LLM。即：

LLM to AI Agent to AGI

What is Agent ?

- Much Like RL (Reinforcement Learning)



What is Agent ?

- 给定特性的目标 AI Agent，Agent 能够自己创建任务、完成任务、创建新任务、重新确定任务列表的优先级、完成新的顶级任务，并循环直到达到目标。
- Agent 让 LLM 具备目标实现能力，并通过自我激励循环来实现给定的目标。即：Agent = LLM + Planning + Feedback + Tool use。

What is Agent ?

- Agent 让 LLM 具备目标实现能力，并通过自我激励循环来实现给定的目标。即：

Agent = LLM + Planning 计划+ Tool use 执行 + Feedback 纠正偏差

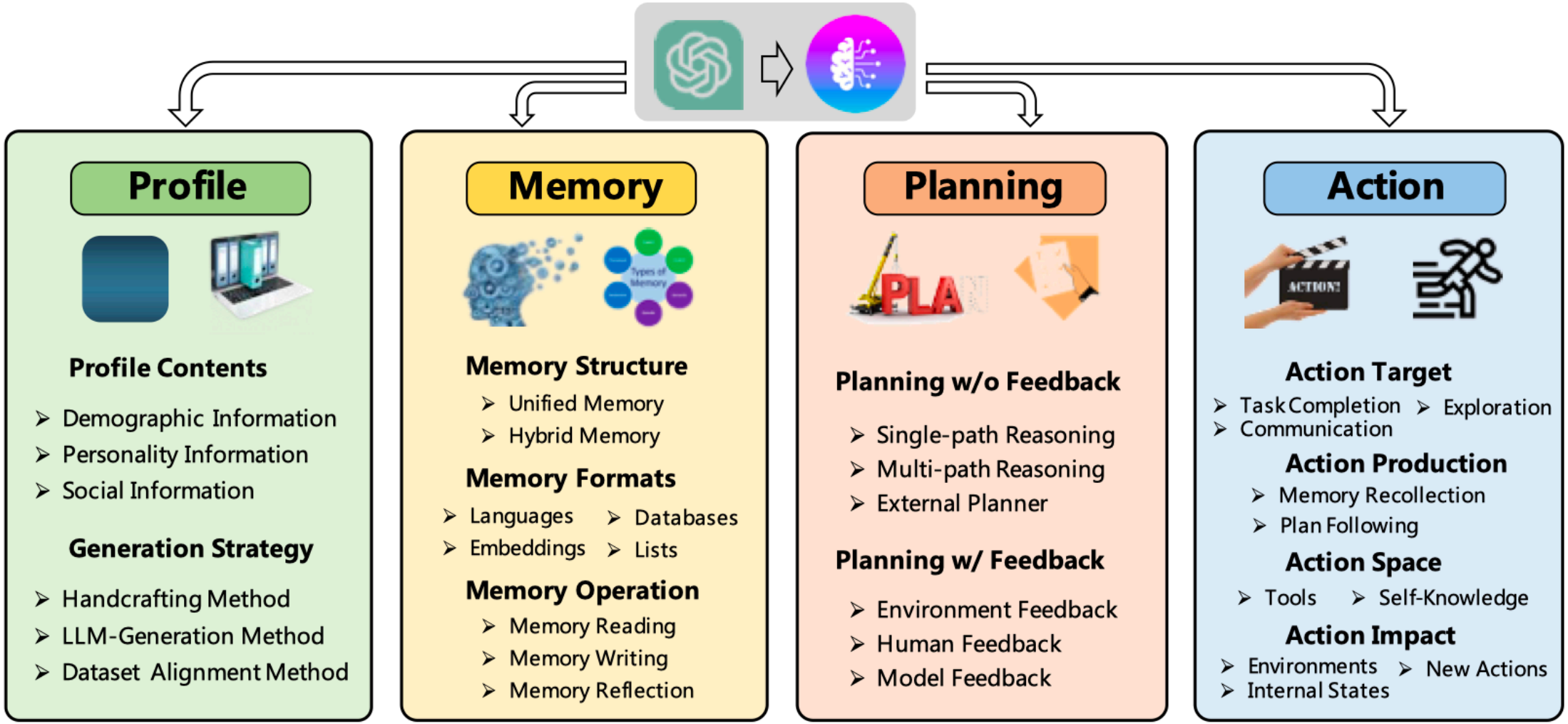


Figure 2: A unified framework for the architecture design of LLM-based autonomous agent.

从 LLM 到 LAM

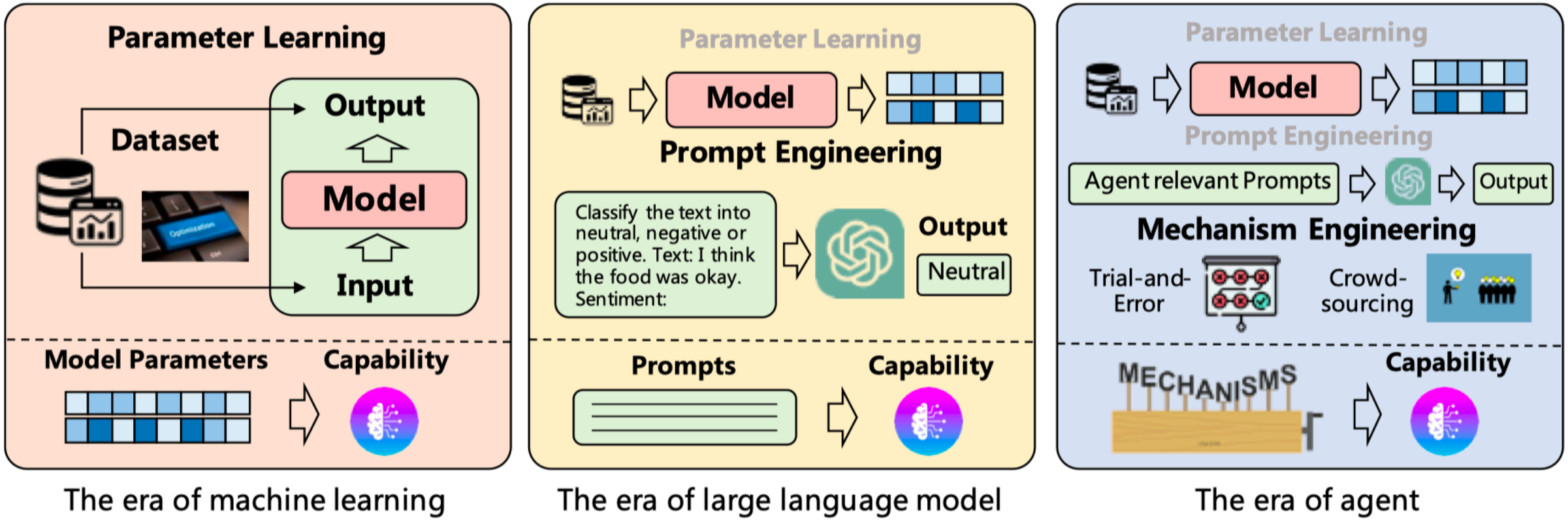


Figure 4: Illustration of transitions in strategies for acquiring model capabilities.

2. 多模态 & Agent

AGI的未来？

1. 人类通过与三维的这是世界交互才能学到 。不一定需要 AI Agent 具备人形，但一定会通过多模态能力来感知、理解和自主探索世界。

——《西部世界》

1. 越来越多的科学家认为 多模态智能体将是 AI 的未来。

——《Bahrckd ydhcc》

多模态

- 目前多模态模型结构大同小异，以 LLM 作为为核心，在多模态输入和输出侧分别加上 encoder 和 diffusion 生成模型。
- Encoder 把图片、音频和视频编码 LLM 所能理解的向量；Diffusion 根据 LLM 输出，生成图片、音频和视频。

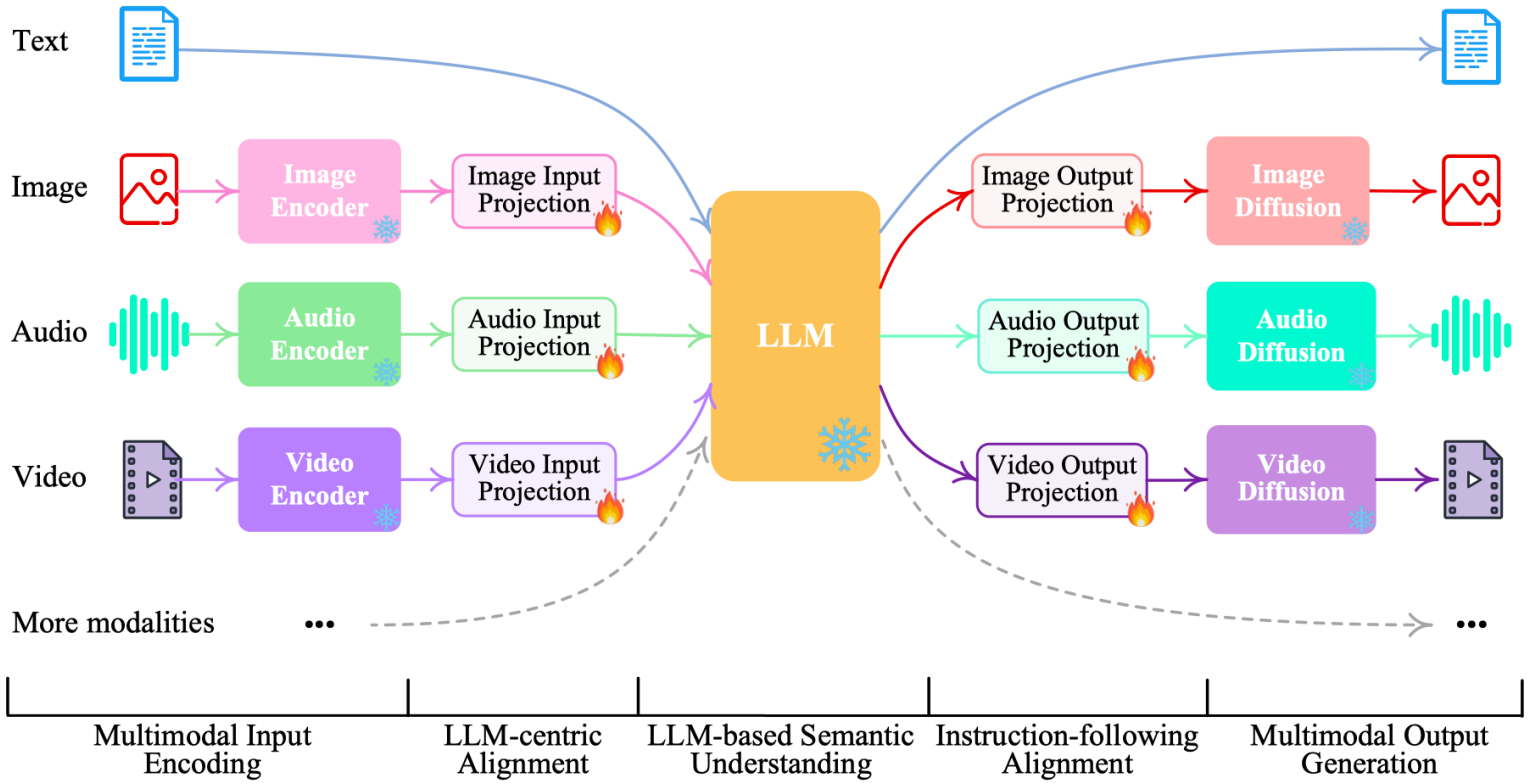
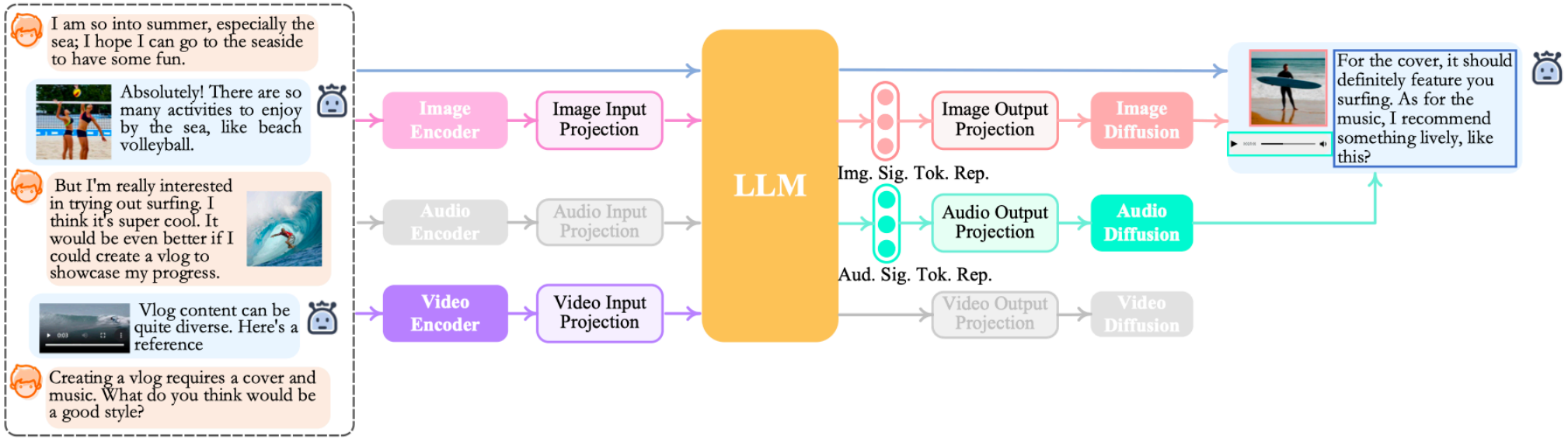


Figure 1: By connecting LLM with multimodal adaptors and diffusion decoders, NExT-GPT achieves universal multimodal understanding and any-to-any modality input and output.

多模态训练过程

- LLM 和 Encoder 间训练 Projection Layer，作为图像、音频和视频输入到 LLM 之间的映射关系；
- LLM 和 Diffusion 间训练 Projection Layer，作为 LLM 输出到图像、音频和视频输出之间的映射关系。
- LLM 外挂 LoRA 用来实现指令微调 Instruction Tuning，把多模态的输入输出数据输入，让 LLM 学会在多模态间进行转换（如输入图片和文字描述的问题，输出文字回复）。



Question ?

1. 为什么 LLM 大模型火起来，甚至众多科学家想通过 LLM + Agent 走向AIG。而多模态大模型的实际效果不好呢？

多模态大模型的未来

- 技术上靠谱多模态大模型很有可能是类 Next-GPT 结构，但其训练方式绝对不是 Instruction Tuning，而是在预训练阶段使用大量的图片、语音、文字甚至视频的多模态语料进行端到端的训练。

Dataset	Data Source	In→Out Modality	Approach	Multi-turn Reason	#Img/Vid/Aud	#Dialog Turn.	#Instance
► *Existing data*							
MiniGPT-4 [109]	CC [10], CC3M [71]	T+I→T	Auto	✗	134M/-/-	1	5K
StableLLaVA [47]	SD [68]	T+I→T	Auto+Manu.	✗	126K/-/-	1	126K
LLaVA [104]	COCO [50]	T+I→T	Auto	✓	81K/-/-	2.29	150K
SVIT [106]	MS-COCO [50], VG [41]	T+I→T	Auto	✓	108K/-/-	5	3.2M
LLaVAR [104]	COCO [50], CC3M [71], LAION [70]	T+I→T	LLaVA+Auto	✓	20K/-/-	2.27	174K
VideoChat [44]	WebVid [5]	T+V→T	Auto	✓	-/8K/-	1.82	11K
Video-ChatGPT [54]	ActivityNet [28]	T+V→T	Inherit	✗	-/100K/-	1	100K
Video-LLaMA [103]	MiniGPT-4, LLaVA, VideoChat	T+I/V→T	Auto	✓	81K/8K/-	2.22	171K
InstructBLIP [15]	Multiple	T+I/V→T	Auto	✗	-	-	~ 1.6M
MIMIC-IT [42]	Multiple	T+I/V→T	Auto	✗	8.1M/502K/-	1	2.8M
PandaGPT [77]	MiniGPT-4, LLaVA	T+I→T	Inherit	✓	81K/-/-	2.29	160K
MGVLID [107]	Multiple	T+I+B→T	Auto+Manu.	✗	108K/-/-	-	108K
M^3IT [45]	Multiple	T+I/V/B→T	Auto+Manu.	✗	-/-/-	1	2.4M
LAMM [97]	Multiple	T+I+PC→T	Auto+Manu.	✓	91K/-/-	3.27	196k
BuboGPT [108]	Clotho [20], VGGSS [11]	T+A/(I+A)→T	Auto	✗	5k/-/9K	-	9K
mPLUG-DocOwl [96]	Multiple	T+I/Tab/Web→T	Inherit	✗	-	-	-
► *In this work*							
T2M	Webvid [5], CC3M [71], AudioCap [38]	T→T+I/A/V	Auto	✗	4.9K/4.9K/4.9K	1	14.7K
MosIT	Youtube, Google, Flickr, Midjourney, etc.	T+I+A+V→T+I+A+V	Auto+Manu.	✓	4K/4K/4K	4.8	5K

Table 2: Summary and comparison of existing datasets for multimodal instruction tuning. T: text, I: image, V: video, A: audio, B: bounding box, PC: point cloud, Tab: table, Web: web page.

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