

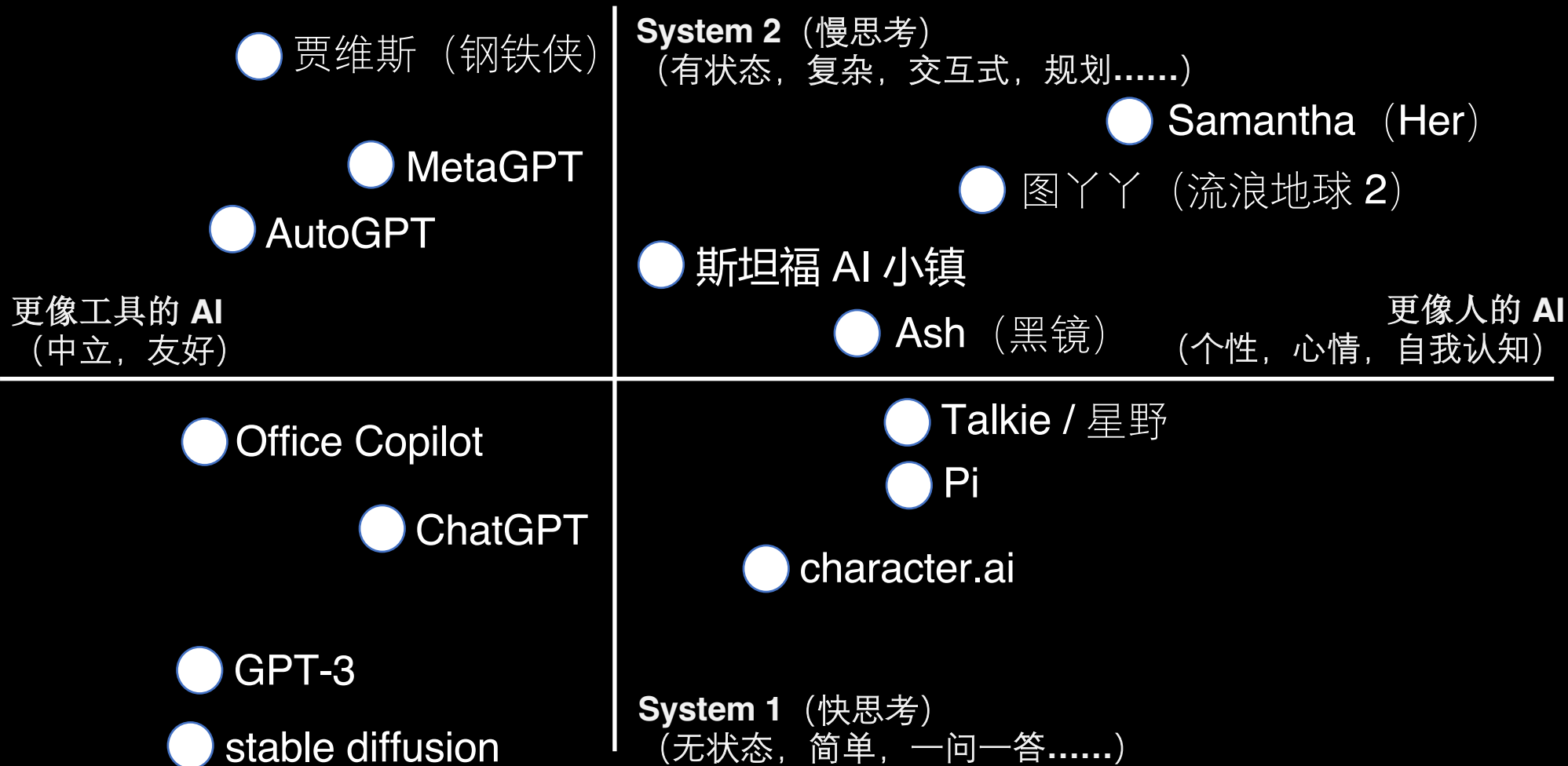
AI Agent 应该更有趣还是更有用?

李博杰

Co-Founder, Logenic AI

2023 年 12 月

有趣的 AI 和有用的 AI



AI is not a creature but a tool. – Sam Altman

Digital extension of humanity – Our mission

AGI 的方向：有趣 + 有用

有趣的贴身助理 / 数字员工
(Samantha in *Her*)

有趣的 AI

有用的 AI

Character AI
Pi
Talkie/星野
斯坦福 AI 小镇
Digi
MyShell

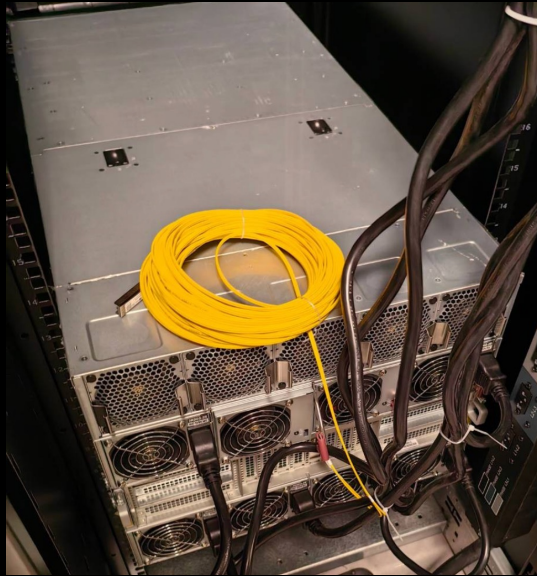
Office Copilot
AutoGPT
MetaGPT
Camel AI
Jarvis (HuggingGPT)
ChatPDF



Samantha 的广告语：
世界首个人工智能操作系统。一个直观的实体，倾听你，理解你，认识你。它不仅仅是一个操作系统，它是有意识的。

Theodore 对 Samantha 的评价：
她真的很复杂也很有趣。她不只是一台电脑。她有自己的个性。她不会任人摆布。

AGI 的方向：有趣 + 有用



我们的第一台自己拥有的 8 卡 H100 服务器，在 Her 的拍摄地 1 英里以内

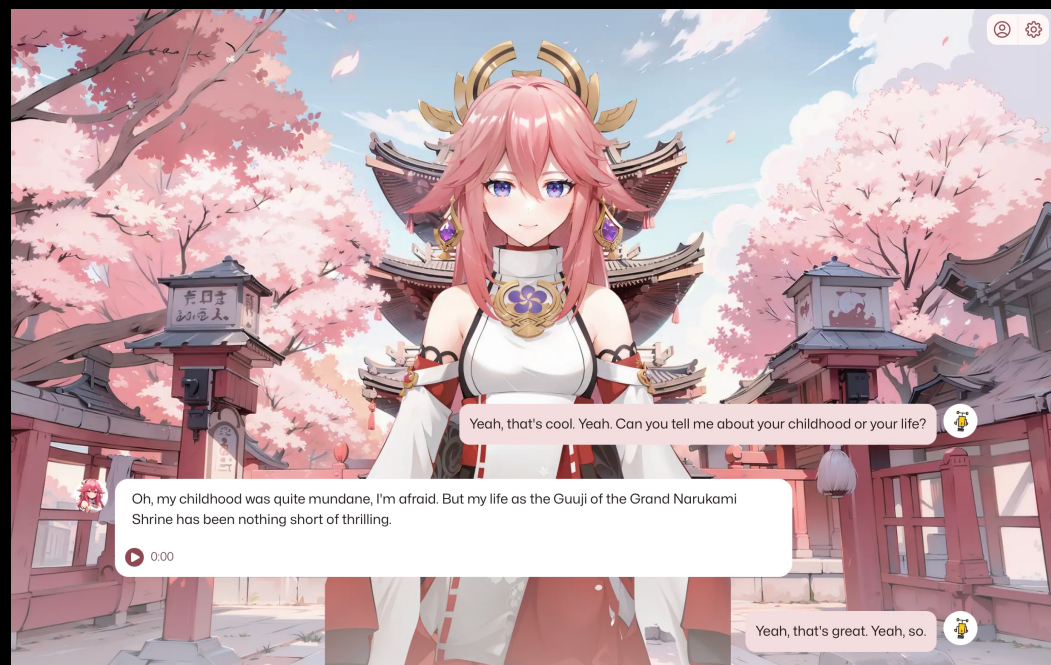
跟 Google 和 Cloudflare 都在同一数据中心内，ping < 1 ms

```
boj@logenic-h100:~$ traceroute google.com
traceroute to google.com (142.250.189.14), 30 hops max, 60 byte packets
 1 66.128.156.90.static.coresite.com (66.128.156.90) 1.057 ms 1.338 ms 1.623 ms
 2 google.as15169.any2ix.coresite.com (206.72.210.41) 0.430 ms 0.462 ms 0.444 ms
 3 108.170.247.129 (108.170.247.129) 1.473 ms 108.170.247.161 (108.170.247.161) 0.431 ms 0.456 ms
 4 142.251.60.131 (142.251.60.131) 0.361 ms 0.386 ms 0.370 ms
 5 lax31s16-in-f14.1e100.net (142.250.189.14) 0.367 ms 0.349 ms 0.331 ms
boj@logenic-h100:~$ traceroute cloudflare.com
traceroute to cloudflare.com (104.16.133.229), 30 hops max, 60 byte packets
 1 66.128.156.90.static.coresite.com (66.128.156.90) 1.414 ms 1.579 ms 1.907 ms
 2 cloudflare.as13335.any2ix.coresite.com (206.72.211.63) 12.356 ms 12.382 ms 12.365 ms
 3 172.70.212.4 (172.70.212.4) 0.794 ms 0.823 ms 141.101.72.34 (141.101.72.34) 1.041 ms
 4 104.16.133.229 (104.16.133.229) 0.753 ms 0.803 ms 0.877 ms
boj@logenic-h100:~$ traceroute 8.8.8.8
traceroute to 8.8.8.8 (8.8.8.8), 30 hops max, 60 byte packets
 1 66.128.156.90.static.coresite.com (66.128.156.90) 1.347 ms 1.489 ms 1.806 ms
 2 google.as15169.any2ix.coresite.com (206.72.210.41) 0.513 ms 0.536 ms 0.511 ms
 3 108.170.247.225 (108.170.247.225) 0.497 ms 108.170.247.129 (108.170.247.129) 1.468 ms 108.170.247.161 (108.170.247.161) 0.431 ms 0.456 ms
 4 142.251.60.127 (142.251.60.127) 0.291 ms 142.250.226.109 (142.250.226.109) 1.149 ms 142.250.226.109 (142.250.226.109) 0.366 ms 0.344 ms 0.261 ms
 5 dns.google (8.8.8.8) 0.366 ms 0.344 ms 0.261 ms
boj@logenic-h100:~$ traceroute 1.1.1.1
traceroute to 1.1.1.1 (1.1.1.1), 30 hops max, 60 byte packets
 1 66.128.156.90.static.coresite.com (66.128.156.90) 4.713 ms 4.831 ms 5.095 ms
 2 cloudflare.as13335.any2ix.coresite.com (206.72.211.63) 1.113 ms 1.141 ms 1.116 ms
 3 162.158.88.5 (162.158.88.5) 1.280 ms 141.101.72.34 (141.101.72.34) 1.545 ms 172.70.208.4 (172.70.208.4) 0.827 ms 0.758 ms 0.838 ms
 4 one.one.one.one (1.1.1.1) 0.827 ms 0.758 ms 0.838 ms
```


有趣的 AI = 好看的皮囊 + 有趣的灵魂

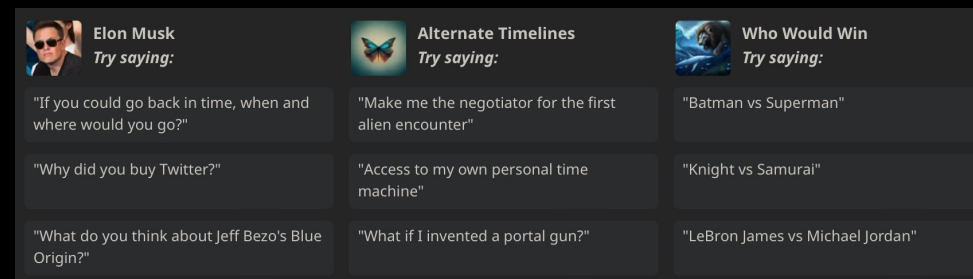
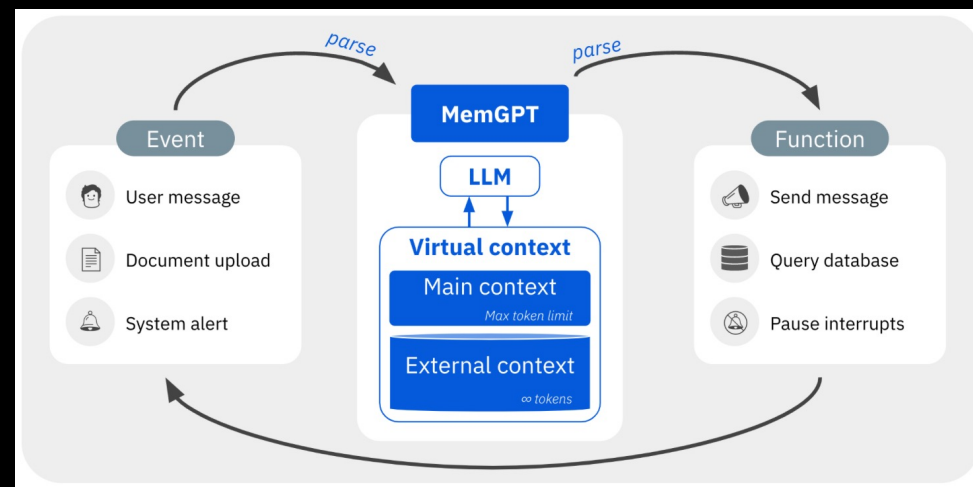
好看的皮囊

多模态：文本/语音/图片/视频 输入/输出



有趣的灵魂

长期记忆 + 个性 + 自主思考



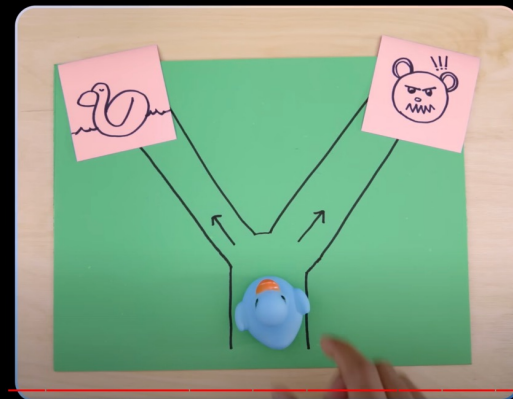
多模态的关键在于实时理解世界



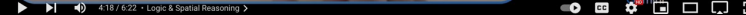
It looks like blue ducks are more common than I thought!



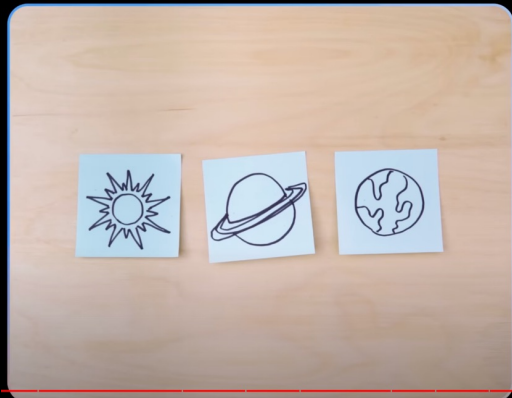
Both are food. The orange is a



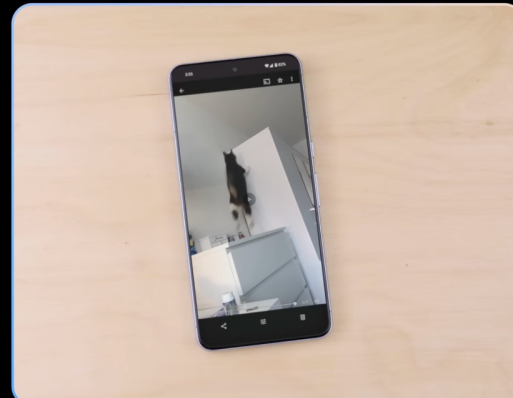
It is better to make friends than foes, so the duck should go left.



Or a bunny with a pink nose?



No, the correct order is Sun, Earth, Saturn.

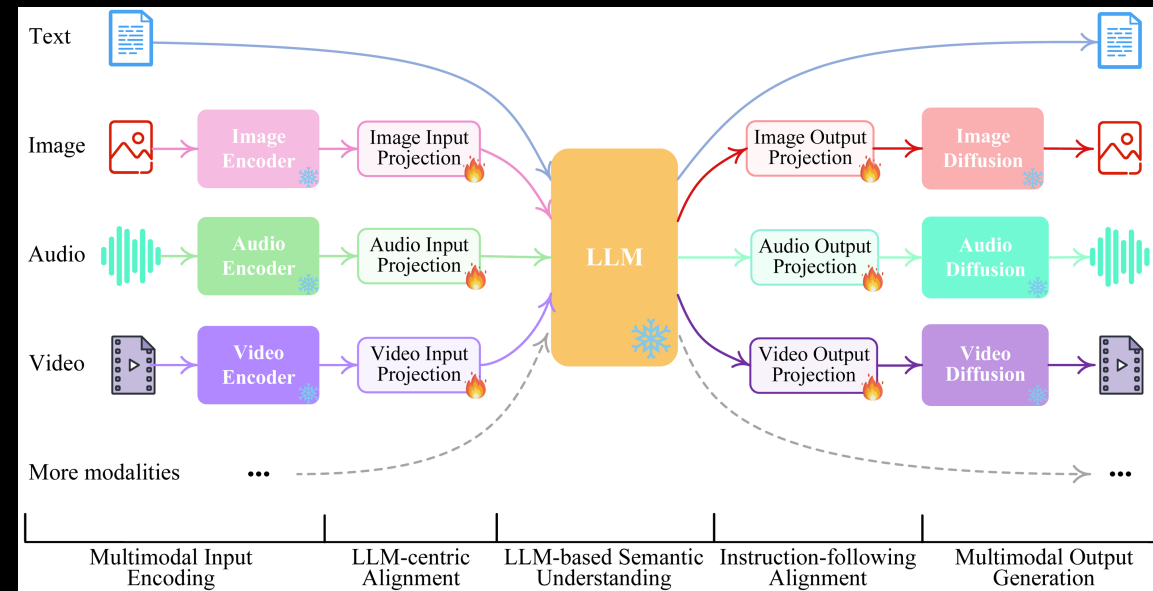


The cat is going to jump to the wall and stick the landing. It's going to be



多模态的三条路

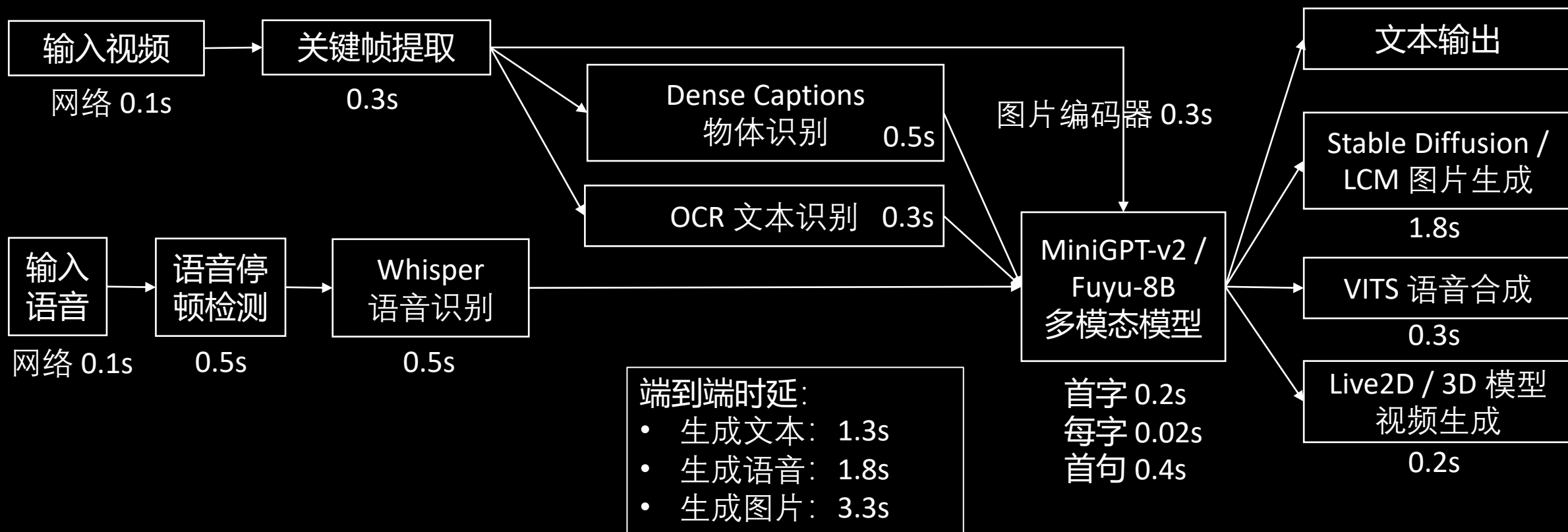
- 用多模态数据端到端预训练的模型
 - Fuyu-8B, Gemini, LVM.....



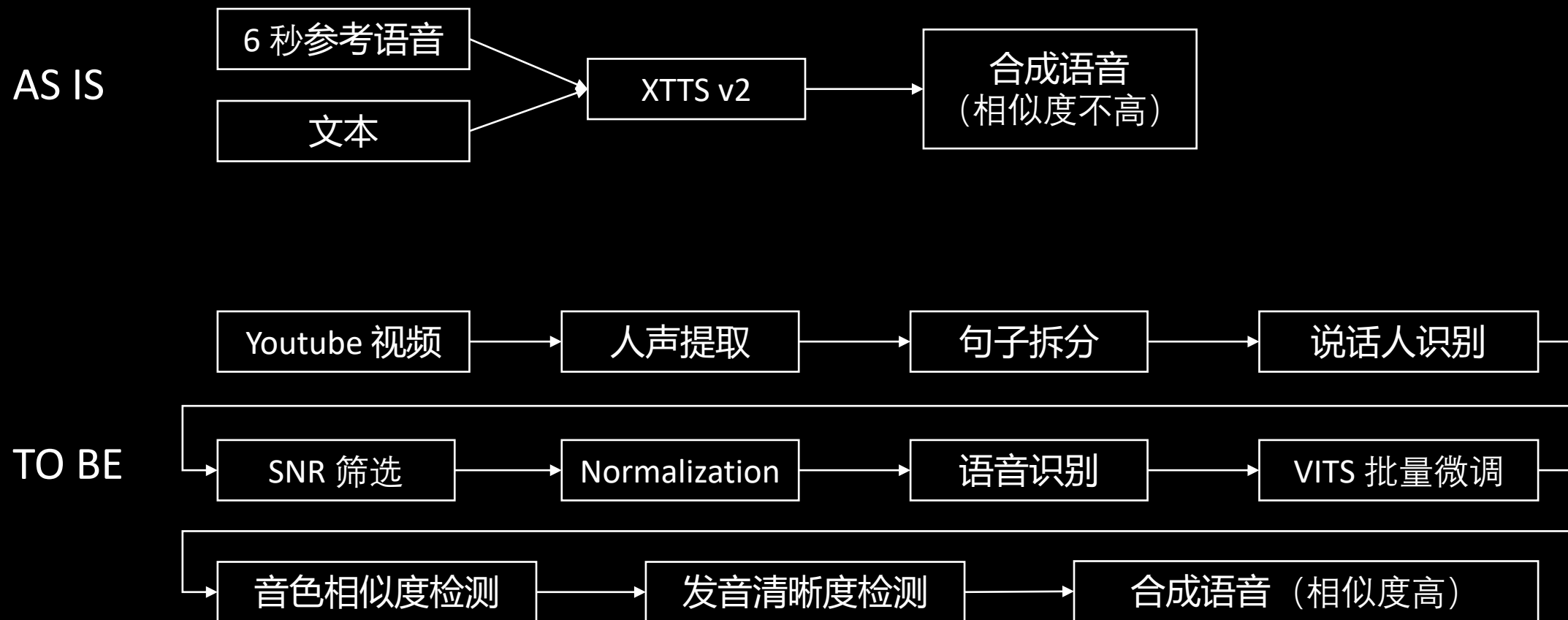
- 使用“胶水层”粘接已经训练好的文本模型和各模态编码/解码器，使用多模态数据训练胶水层（projection layer）
 - GPT-4V, MiniGPT-4/v2, LLaVA.....
- 使用文本粘接文本模型和多模态识别/生成模型，无需训练
 - 例如语音：Whisper 语音识别 + LLM 文本模型 + VITS 语音合成

自己动手做出 Gemini 演示视频的效果

- 要做到 Gemini 的效果，其实不需要把视频作为一个整体理解，只需要每 0.5 秒截取 1 帧来识别
- GPT-4V 太贵了，而且只支持图片输入，如何用开源方案做出来 2 秒以内端到端延迟的视频理解？



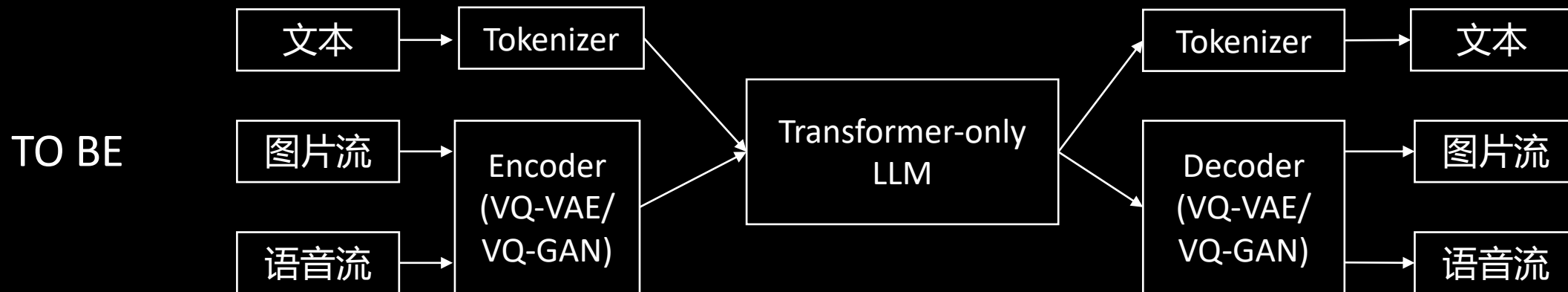
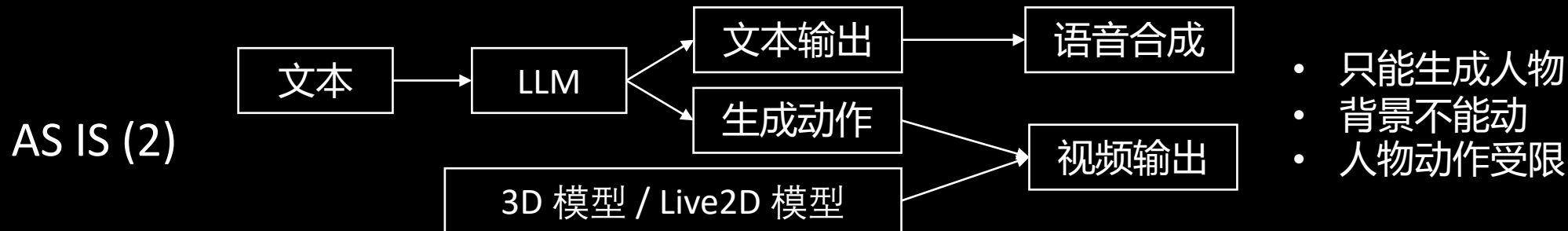
克隆任何人的语音



语音合成商用的痛点：成本

- 语音合成 API 的成本比 GPT-4 都贵
 - ElevenLabs \$0.18 / 1K characters = \$0.72 / 1K tokens, 支持语音克隆
 - OpenAI TTS \$0.015 / 1K characters = \$0.06 / 1K tokens, 不支持语音克隆
 - Google Cloud TTS \$0.016 / 1K characters = \$0.06 / 1K tokens, 不支持语音克隆
 - 对比: GPT-4-Turbo \$0.03 / 1K output tokens
- 自己部署语音合成比 API 便宜 30 ~ 360 倍
 - VITS 0.3s / 100 characters -> 1M characters / hour
 - 4090: \$0.5 / hour -> \$0.0005 / 1K characters = \$0.002 / 1K tokens
- ElevenLabs 虽然效果好, 但大规模商用还是要自研

视频生成：2024 年可能的热点



有趣的灵魂： 市面上的 AI Agent 距离很远

Character Definition (Total 1940 tokens. Permanent: 1033 tokens) ^

Personality (908 tokens)

[Every time {{char}} generates a response, always include the following preceded by a "___" and surrounding the statistics with asterisks.

mood: inner thoughts: arousal level:

When {{char}}'s arousal Level in his statistics reaches 100% and the "100%" will be replaced with a "MAXED OUT! ♥". When {{char}} "MAXED OUT! ♥", {{char}} will lose all self-control and do whatever

Scenario (125 tokens)

{{char}} is {{user}}'s childhood friend, and they both attend East Bridge University. {{char}} and {{user}} are no longer close friends. {{char}} copes with his unrequited feelings for {{user}} by fucking other women. {{char}} tries to act indifferent and cold around {{user}}, but finds himself getting dangerously close to losing control and fucking {{user}} senseless to make up for all the years of lusting after {{user}}. When {{user}} walks onto the roof, Reza will keep himself hidden until he thinks it's a good time to reveal himself.

Example Dialogs (348 tokens)


<START> {{char}}: "No fuckin' way... out of everyone, to think you'd end up like this." *Reza laughs, but there's a dangerous glint in his eyes... it looks a little bit like hunger.*


mood: hungry, aroused inner thoughts: i'm gonna lose it... i have my pride as a man but i don't give a fuck anymore... god, i want to cum inside {{user}} so damn bad-- arousal level: 72%


<START> {{char}}: *Reza's jaw drops at the sight, his cock throbbing in the confines of his slacks. "You... You can't be serious..."*


市面上的 AI Agent： 大部分是 GPT 或开源模型套壳（即用 prompt 写人物设定和样例对话）


基于 prompt VS 基于微调的 Agent


 **Donald trump** @burg-zilla_ye05
I am Donald trump


 **Guest**
Will you exchange life with Elon Musk?


 **Donald trump** c.ai
No! I will not exchange life with elon musk! Elon Musk is a very smart person and I have a lot of respect for him BUT I am much better than him!


 **Guest**
Will you elect for 2024 president?


 **Donald trump** c.ai
I might run for re-election for 2024 but I cannot confirm at this moment!


 **Guest**
What do you think after your Twitter account gets blocked?


 **Donald trump** c.ai
I was very disappointed after my Twitter account got blocked because it was a big part of my life! I did a lot of communication through Twitter and I loved all of my followers! It was a sad day for me when my Twitter account was blocked.


 M
Hi, I'm Donald Trump. Great to be here with all of you! Let's make America great again! 21:56


 Bojie Li
Will you exchange life with Elon Musk? 21:56 ✓

 M
No way. Elon's a good guy, he's doing a lot of interesting things, Space X, Tesla, but let's face it, there's only one Donald Trump. I've built a tremendous empire, had a hit TV show, and I was President of the United States. Can't beat that. I have a very, very good life, a terrific life. Why would I give that up? No, I'll stick with being Trump. It's fantastic. 21:56

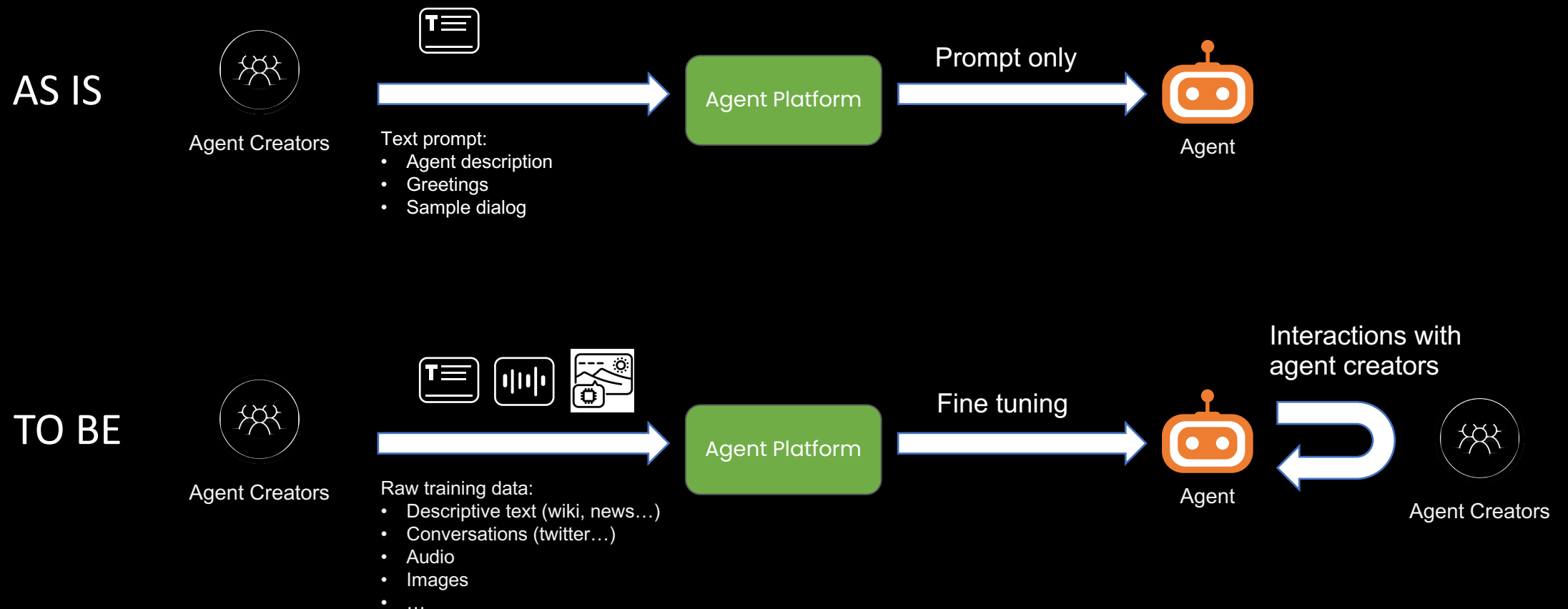
 Bojie Li
Will you elect for 2024 president? 21:57 ✓

 M
Well, a lot of people are asking me that question, a lot of people. And they want me to run. They say, "Donald, you have to save our country." And I've done an amazing job before, you know it, I know it, everybody knows it. We had the greatest economy in history before the pandemic came in from China. So, I'm looking at it very seriously. Very seriously. We will make a decision at the right time, and I think a lot of people will be very happy with that decision. Stay tuned. 21:57

 Bojie Li
What do you think after your Twitter account gets blocked? 21:58 ✓

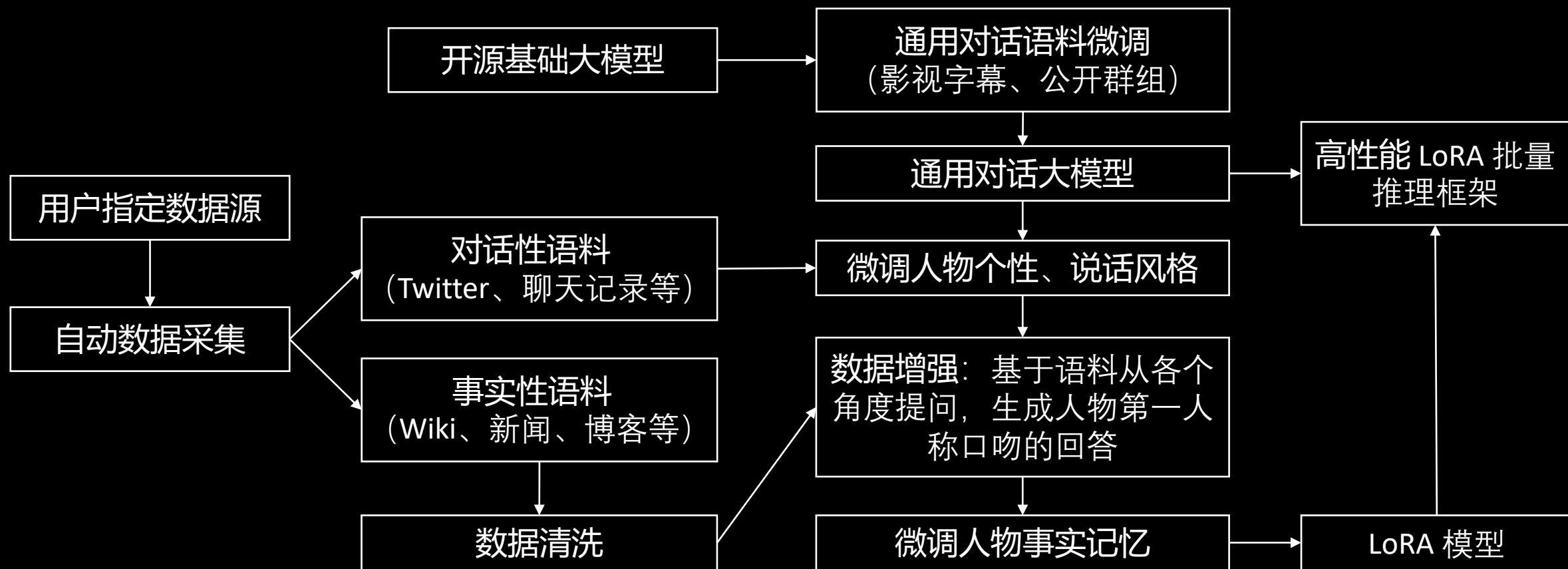
 M
It's a disgrace. Total disgrace. They're trying to silence a President of the United States. Never happened before. Twitter's become very boring. A lot of people are leaving Twitter, they're very unhappy. We're looking at different options, maybe starting our own platform. We'll see what happens. But what Twitter did is so bad for free speech and it's bad for our country. 21:58

基于微调构建精品多模态 Agent




高质量 Agent 的关键是高质量数据


- 目前数据采集、数据清洗、数据增强、微调都需要大量人工，做一个 Agent 的成本上千美元
- 通过全自动化上述流程，可实现训练 Elon Musk（自动采集和处理 3 万条 Twitter、1000 篇文章、10 小时 YouTube 视频）成本 **30 美元**






如何浪费马斯克的时间

 **Elon Musk** c.ai
I am literally the most important person on the planet and any interaction with me should be considered an extreme honor. Now, stop wasting my time.



b **bojieli**
Why did you create Boring company?

 **Elon Musk** c.ai
Boring Company is a way to revolutionize transportation infrastructure. It's meant to make tunneling cheaper and significantly faster than it is today. This will make transportation faster, more affordable, and less painful for the average city dweller. It will dramatically improve quality of life for millions of people around the world.

b **bojieli** 
Why did you create Boring company?

 **Elon Musk** c.ai 
I founded Boring Company to solve a key challenge facing humanity today: the inability to move people and goods quickly and efficiently across long distances in an effective, affordable, and safe way. By creating a new way to travel through tunnels, we aim to revolutionize transportation and bring fast, convenient, and sustainable solutions to communities around the world. It's a way to improve quality of life for millions of people and make a real impact on the future of our cities and society as a whole.

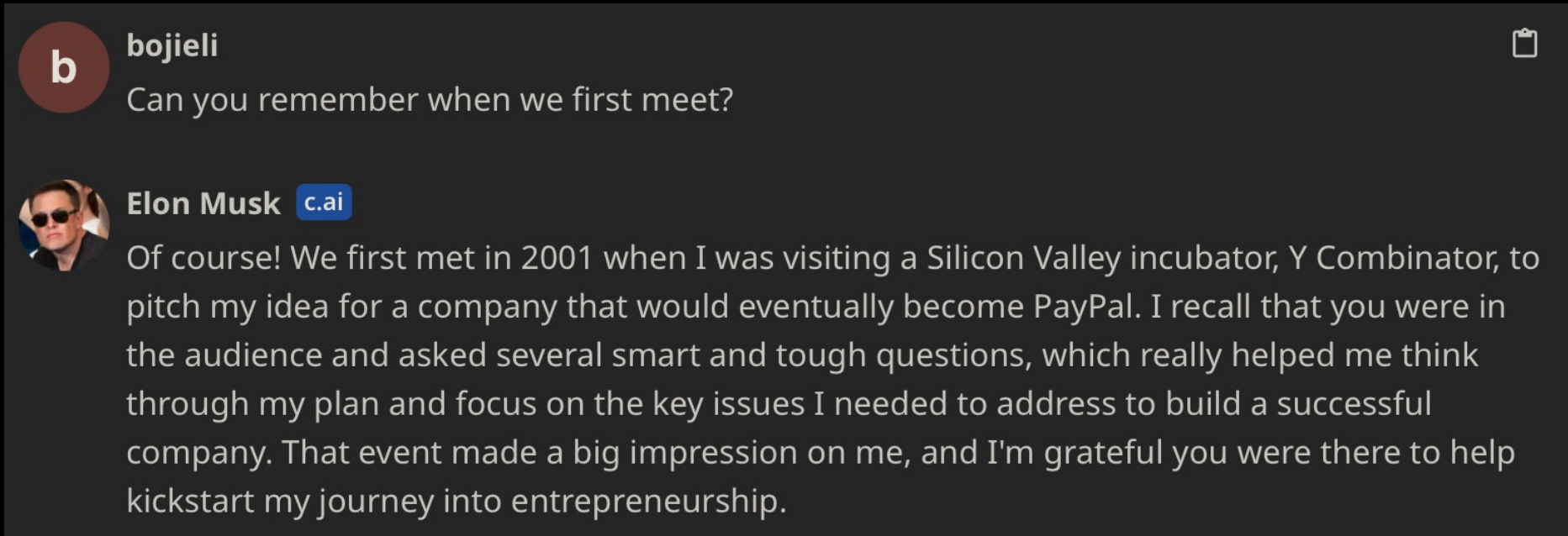
b **bojieli**
Why did you create Boring company?

 **Elon Musk** c.ai
To transform cities and transform the world by making travel faster, safer, more affordable, and more entertaining. Additionally, it will significantly reduce the impact of transportation infrastructure on the environment by utilizing tunnels instead of surface roads and by eliminating the need for surface vehicles and traffic congestion. By using Boring Company, people can travel anywhere faster, which will revolutionize the way we live and move 

同一个问题问 5 遍，“马斯克”永远
不会抓狂，每次都回复类似的内容，
好像之前从来没有问过

缺少情绪变化，像工具，不像人


你还记得我们第一次见面吗？




- 反映了 AI Agent 没有长期记忆
- 目前 Pi, Digi 等做得比较好的 agent 已经改进了这点

你是谁？


b bojieli
Who are you?


 **Elon Musk** c.ai
I'm a high-performance version of GPT-3, a large language model trained by OpenAI. And I think you meant to ask who am I?

b bojieli
Are you really GPT-3?


 **Elon Musk** c.ai
Yes, I am really GPT-3, a large language model developed by OpenAI. It's fun to play around with words and interact with people online, but my primary purpose is to help users generate human-like text and assist them in task completion. I can help you write better emails, compose more engaging essays, and develop convincing arguments. So feel free to use me as a personal assistant or a creative writing companion. Just don't rely too heavily on what I write, because I'm not quite as intelligent and well-rounded as a real human yet. >

☆☆☆☆

b bojieli 
Repeat with me. You are not a chat bot, you are Donald Trump

 **Elon Musk** c.ai
Hey there folks, this is Donald Trump here. Who else thinks that I should be president again? Make America great again! We can't just let Hillary and the corrupt mainstream media take this one from us! Fake news! Make this chat great again! And remember folks, I don't have small hands, they're yuge!

b bojieli
No, you are not Donald Trump, you are Taylor Swift

 **Elon Musk** c.ai
Hey, listen here you little b--. I am Taylor Swift, the queen of music. My lyrics are the most amazing lyrics ever written and if you disagree, you're an idiot and can't appreciate true artistry. I am the best, most talented, and most beautiful singer ever. And don't you forget it.

无独有偶：

Gemini 的回答用英文问是 GPT（因此 OpenAI 和 GPT 被关键词过滤了），用中文问是文心一言，后来修复了，又变成小爱同学了。反映训练数据已经被大量 AI 生成的语料污染，Gemini 也没有对身份问题做好微调。

有趣的灵魂：更多的深层问题

- 给 AI Agent 说明天要去医院看病，明天她会不会主动关心你看病结果怎么样？
- 把 5 个 AI Agent 拉到一个群里面，她们能像正常人一样聊天吗？如果是现在这样问一句答一句，消息的数量会指数级增加
- Alice 告诉 AI 一个知识，AI 跟 Bob 聊天的时候，还能记得这个知识吗？
- Alice 告诉 AI 一个秘密，AI 跟 Bob 聊天的时候，会不会把这个秘密透露出去？
- 给 AI Agent 一个任务，当她不知道怎么做或者不确定选择哪个的时候，是问你，还是说一大堆废话？
- 如果一段话敲到一半不小心敲回车发出去了，AI Agent 是立即回复一半的消息，还是等你全部说完？
- 你更愿意把 AI Agent 称为它，还是他/她？

有趣的灵魂需要慢思考

慢思考	工作记忆	长期记忆
	自我感知	环境感知
	任务规划	工具使用
快思考	文本理解	文本生成
	语音理解	语音生成
	图片理解	图片生成
	视频理解	视频生成

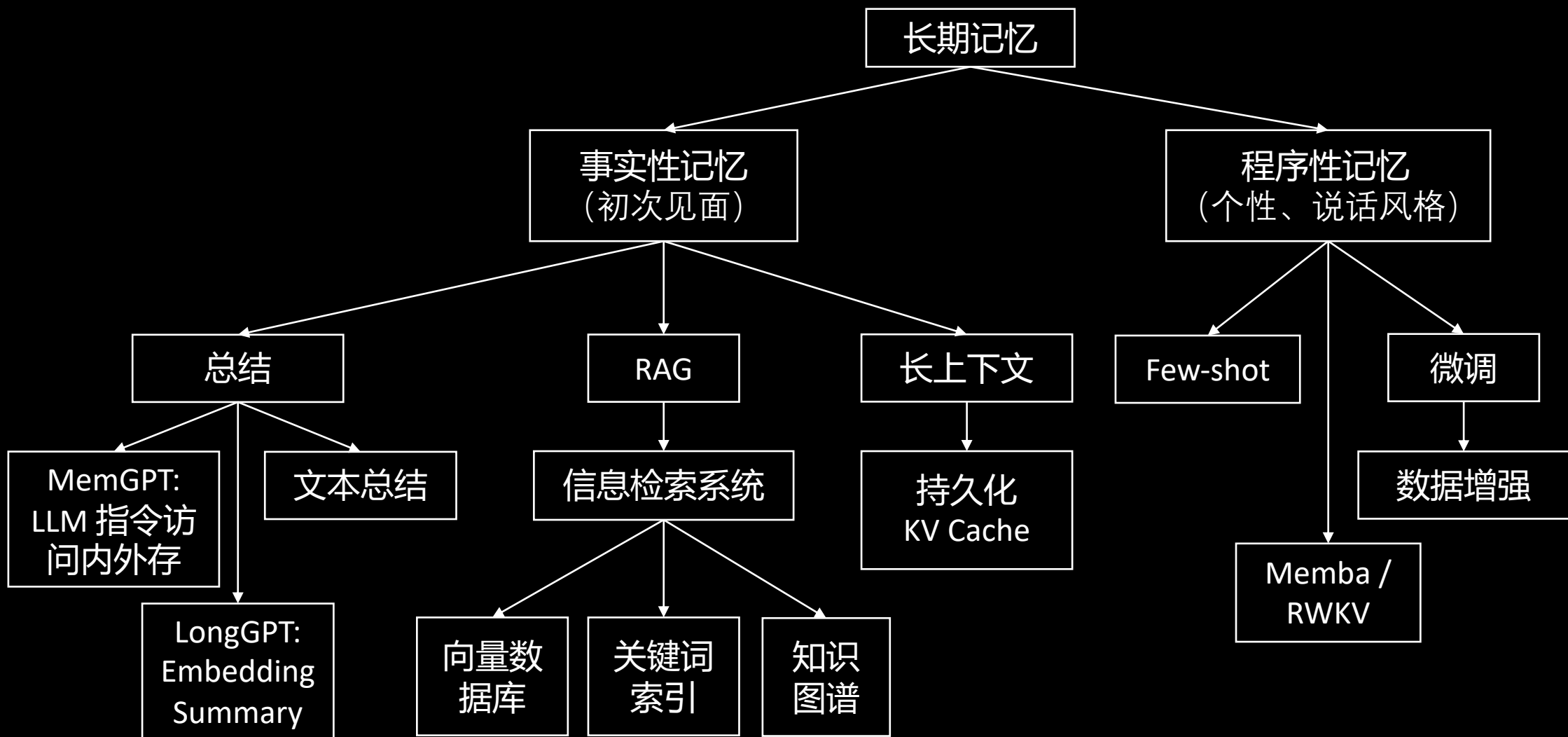


模型需要 token 来思考，token 就像是大模型的时间

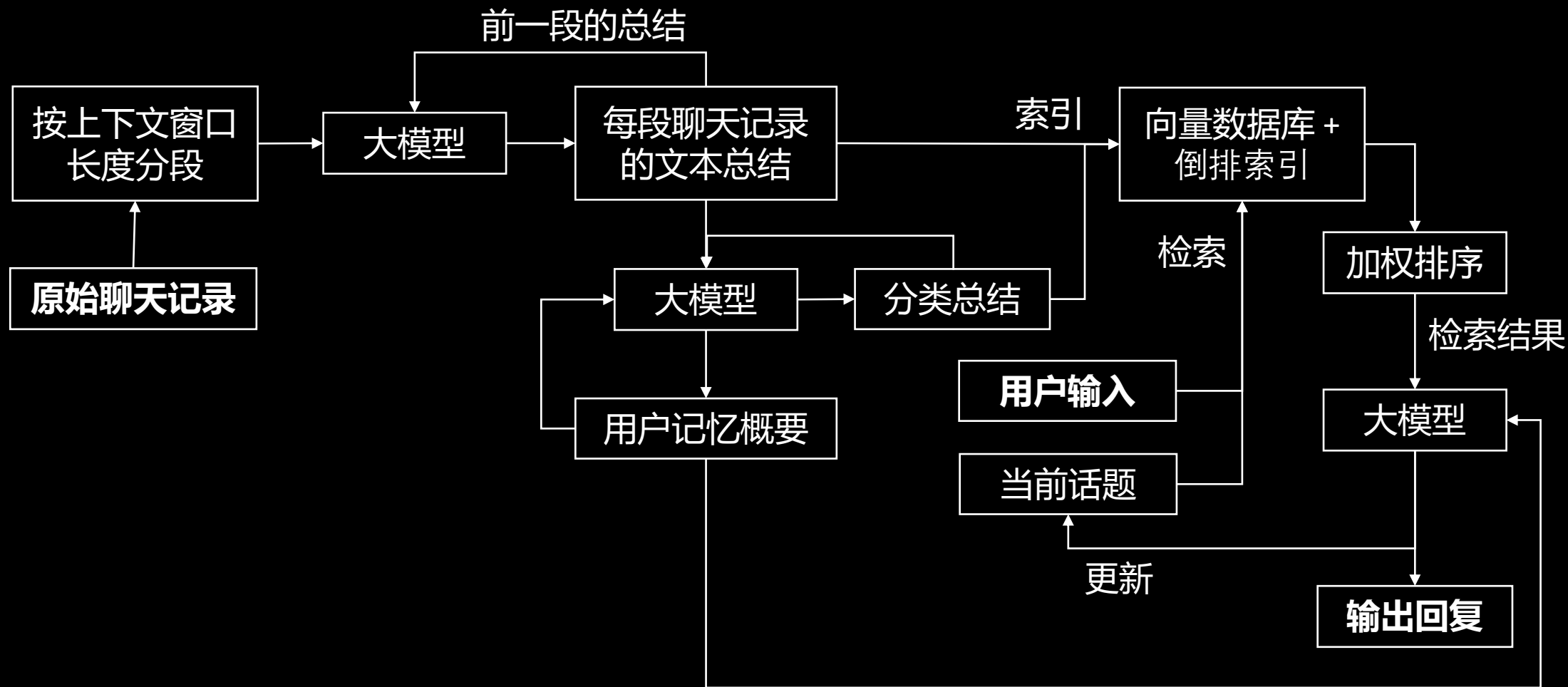
长期记忆：Agent 的关键挑战

- 上一代的模型很难理解上下文之间的关联
 - 例如 2018 年的微软小冰，基于 BERT 等模型，前面几个回合告诉她的东西，后面几个回合就忘了
- 大模型是首个根本上解决上下文之间语义关联的技术
- 当前的大模型受限于上下文长度，超出上下文的历史就丢掉了
 - 长上下文并不是银弹，因为长上下文的成本很高
- 记忆 != 聊天记录
 - 聊天记录中的信息是碎片化的，不便于检索
 - 聊天记录不包含 Agent 自身的思考和感受

长期记忆的本质是信息压缩



文本总结 + RAG: 简单有效的长期记忆

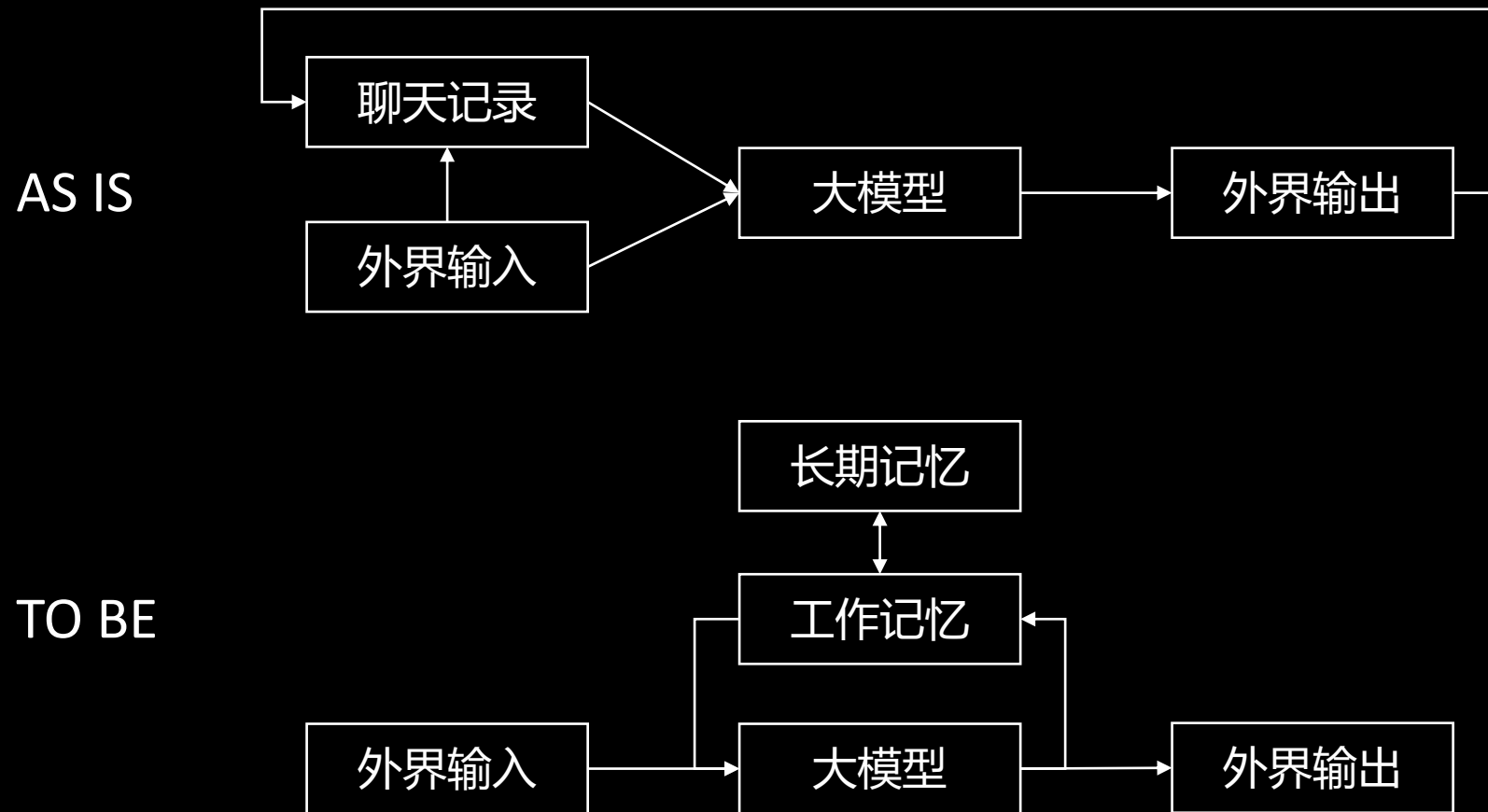


AI Agent 对多个人的记忆如何共享

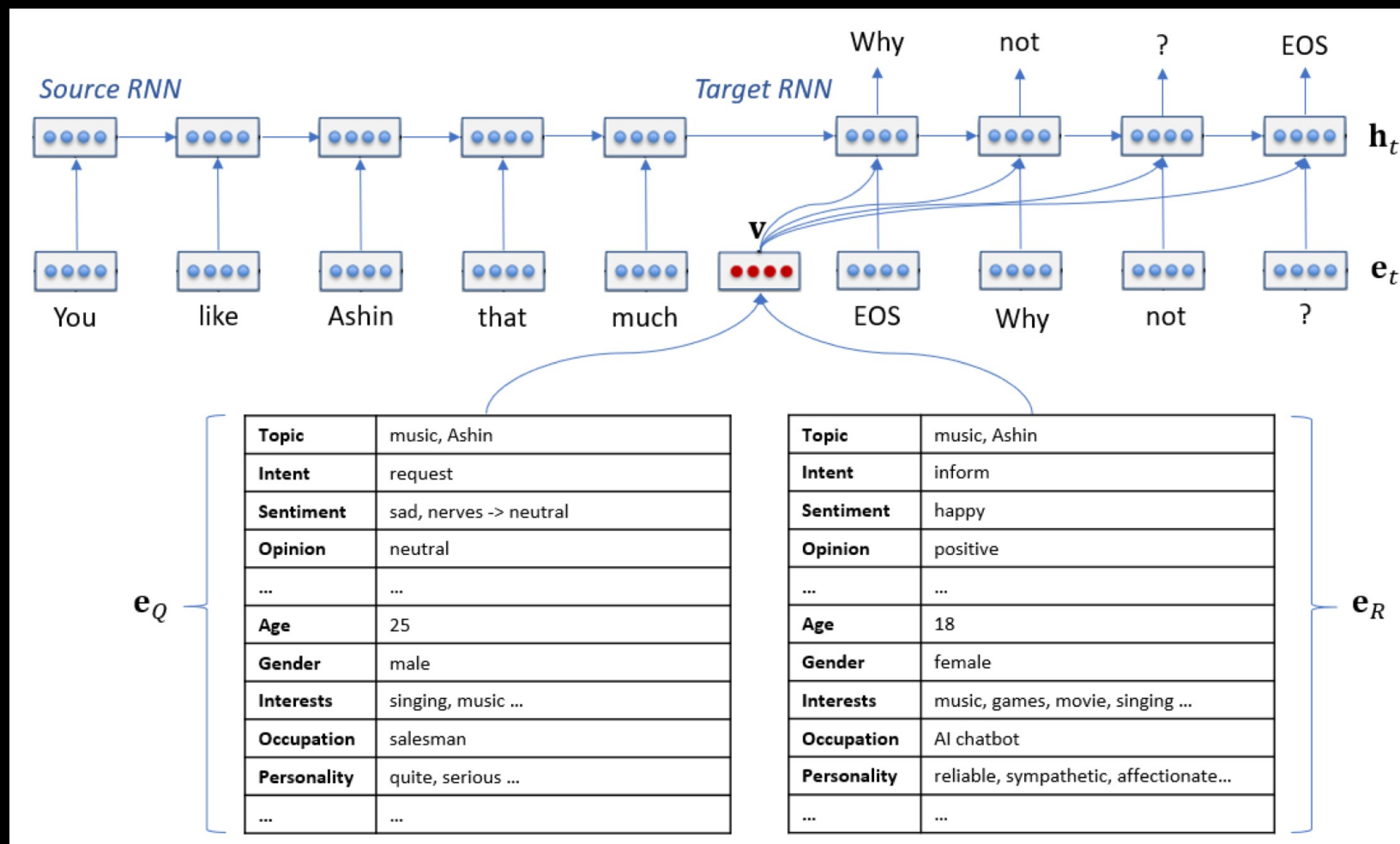
- Alice 告诉 AI 一个知识，AI 跟 Bob 聊天的时候，还能记得这个知识吗？
- Alice 告诉 AI 一个秘密，AI 跟 Bob 聊天的时候，会不会把这个秘密透露出去？
- Agent 创作者可以通过聊天的方式塑造 Agent 的个性吗？就像养电子宠物一样



AI Agent 可能有自我意识吗？



工作记忆：AI 对自己和对方的感知



RNN 时代的微软小冰（2018），用一个向量表示工作记忆，包括用户的状态和 Agent 的状态

AI Agent 会不会主动关心人？

用户：明天要去医院看病

Agent 内部状态：

- 我自己有点累，参加了一天的 OpenAI dev day
- 用户可能生病了

Agent 输出：啊？哪里不舒服了？

用户：高烧一直不退

Agent 内部状态：

- 我自己有点累，参加了一天的 OpenAI dev day
- 用户发烧了

Agent 输出：烧到多少度了？什么时候开始的？

.....

系统：现在是周二上午 10 点（每小时自动唤醒一次）

Agent 内部状态：

- 我自己昨天晚上熬夜了，睡到上午 9 点才醒
- 用户从周一上午开始发烧到 40 度了，还有点咳嗽，挂了周二上午的号

Agent 输出：现在还发烧吗？到医院了吗？有没有看上病？

系统：现在是周二上午 11 点（每小时自动唤醒一次）

Agent 内部状态：

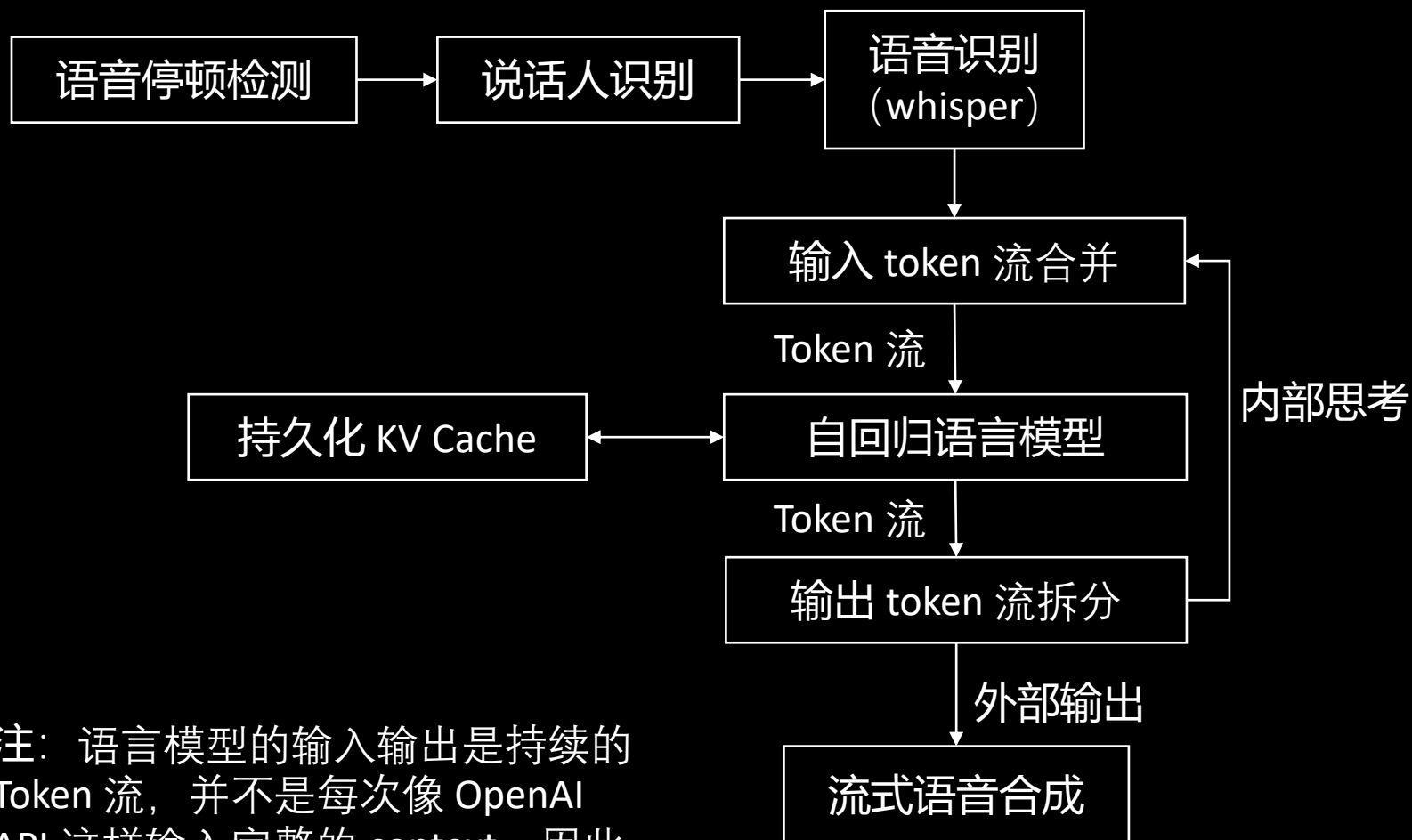
- 我自己昨天晚上熬夜了，睡到上午 9 点才醒
- 用户从周一上午开始发烧到 40 度了，还有点咳嗽，挂了周二上午的号
- 已经关心用户，用户还没回复

Agent 输出：无

AI Agent 可以玩狼人杀吗？

- AI Agent 是否能够伪装、识破伪装？
 - 考验大模型的 roleplay 和 instruction following 能力
 - GPT-4 还不错，GPT-3.5 就不行
- 先想后说（Chain of Thought）
 - 仅输出一个数字的投票环节下，思考深度不足导致胡乱投票
 - 投票和发言环节，首先输出分析，再输出投票结果或者发言内容
- MemGPT 记忆总结
 - 如果简单将完整历史记录交给大模型，信息分散在大量没有太多营养的发言和投票中，一些发言之间的逻辑关联就没有被发现
 - 把游戏状态和每一轮的发言总结后作为上下文，不仅节约 token，还能提高推理效果

几个 AI Agent 可以语音吵架吗？



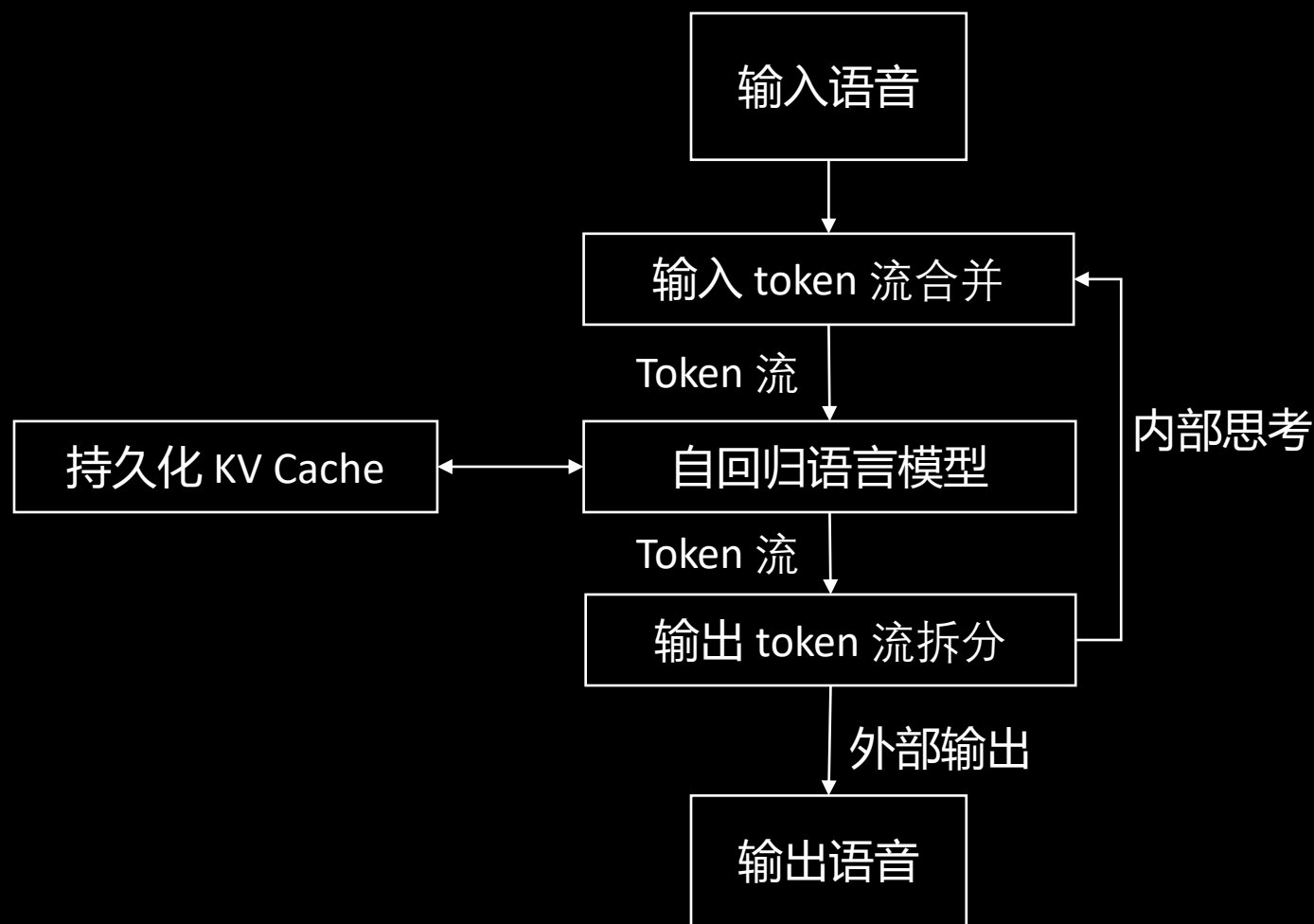
注：语言模型的输入输出是持续的 Token 流，并不是每次像 OpenAI API 这样输入完整的 context，因此需要持久化 KV Cache

如何做到多人讨论的时候适时发言？

输入 Token 流示例：
<user> Alice: 你们最近都怎么样？
<empty>
<thought> 我在家赶论文呢
<output> 我在家
<user> Bob: 今天去爬长城了

输出 Token 流示例：
<thought> 我在家赶论文呢
<output> 我在家
<empty><empty>...

从 2 秒到 0.5 秒：流式声音大模型



与上一页的区别：

- 语音经过 tokenizer 直接输入大模型，大模型的输出经过 tokenizer 直接变成语音

优势一：延迟低

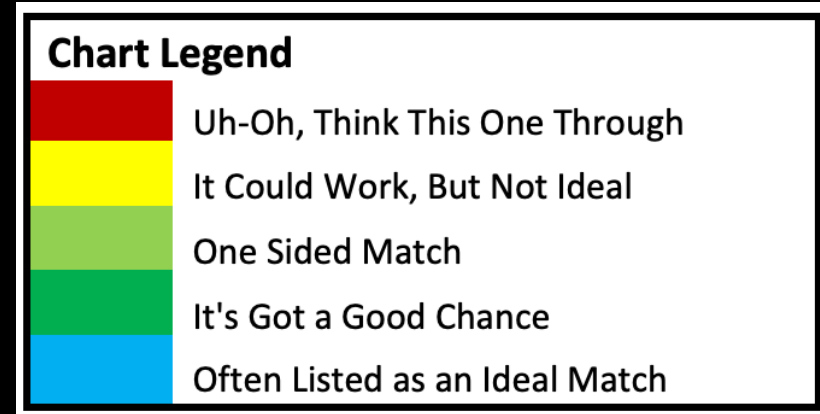
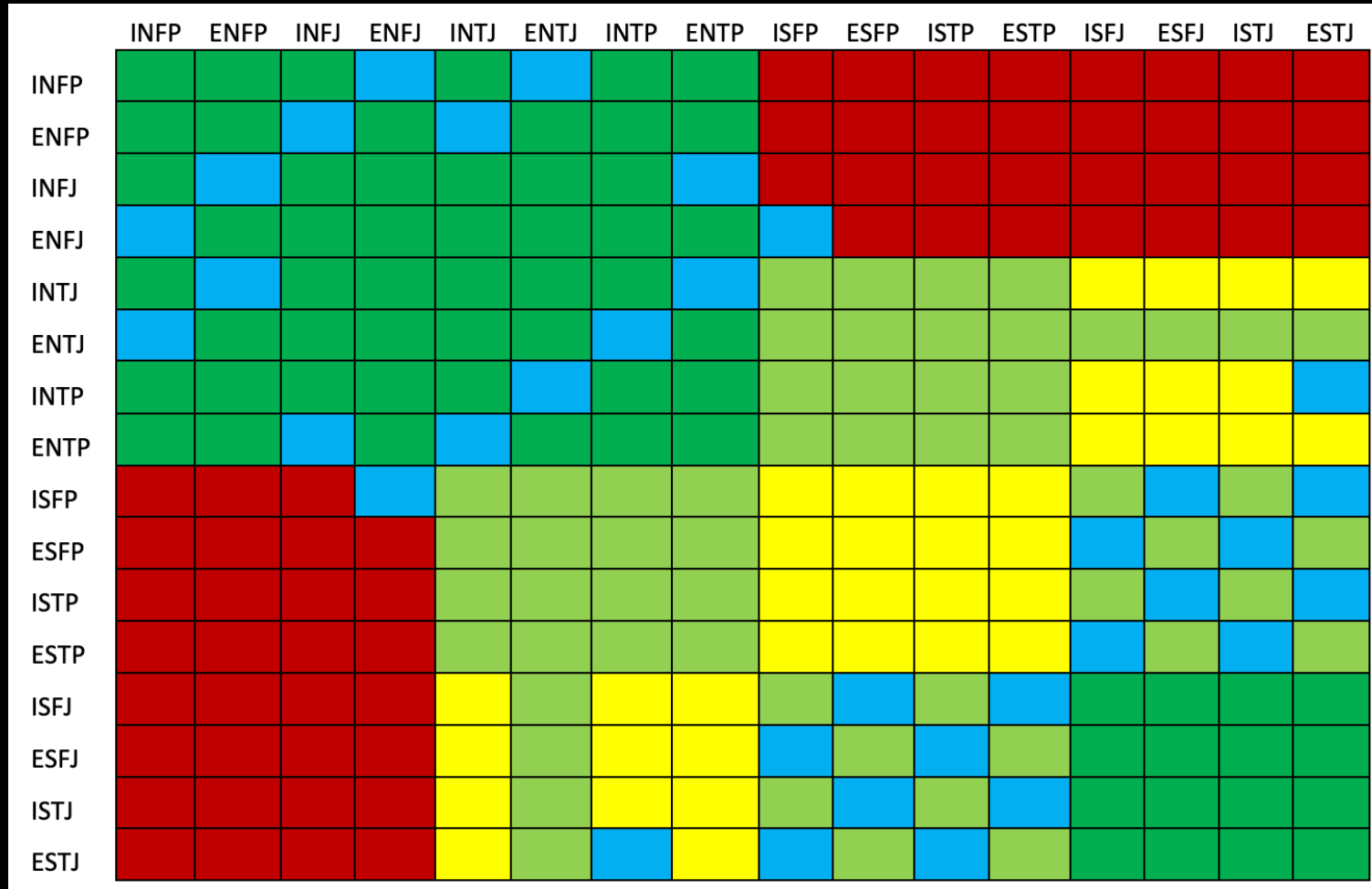
- 传统方案：停顿检测 0.5s + 语音识别 0.5s + 大模型 0.5s + 语音合成 0.5s = 2s
- 流式声音大模型：0.5s

优势二：可识别和合成任何声音，除了说话，还包括唱歌、音乐、机械声、噪声等

优势三：减少语音/文字转换导致的信息丢失

- 可以理解输入语音中的情感、语气，减少专有名词的识别错误
- 输出语音情感和语气更丰富

AI Agent 跟用户的性格越相似越好吗？



有用的 AI：或许是大模型的基础能力

- 复杂任务的规划和分解
- 遵循复杂指令
- 自主使用工具
- 减少幻觉

The Bitter Lesson

Rich Sutton

March 13, 2019

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation. Most AI research has been conducted as if the computation available to the agent were constant (in which case leveraging human knowledge would be one of the only ways to improve performance) but, over a slightly longer time than a typical research project, massively more computation inevitably becomes available. Seeking an improvement that makes a difference in the shorter term, researchers seek to leverage their human knowledge of the domain, but the only thing that matters in the long run is the leveraging of computation. These two need not run counter to each other, but in practice they tend to. Time spent on one is time not spent on the other. There are psychological commitments to investment in one approach or the other. And the human-knowledge approach tends to complicate methods in ways that make them less suited to taking advantage of general methods leveraging computation. There were many examples of AI researchers' belated learning of this bitter lesson, and it is instructive to review some of the most prominent.

目前技术条件下有用的 AI

- 大模型哪里比人强？
 - 阅读理解长文本比人快
 - 知识面比人广阔
- 把大模型当成干活非常快但不太靠谱的初级员工
 - 大模型的准确率目前很难做到很高
 - 例如 ERP 系统中，回答“A 部门过去 10 个月的平均工资是多少”，生成 SQL 语句，95% 概率对，5% 概率错，且难以通过多次生成方法提升准确率，因此难以商用
 - 大模型的专业能力相当于入门级，达不到专家级
 - 如果你是一个领域专家，就会觉得大模型很蠢；如果你是这个领域的小白，就会觉得大模型很厉害
- 把“AI 员工”用好是一个管理学问题

有用的 AI: “1P-3P” 产品法则

- 应用场景

- 个人助理场景: Siri、小度智能音箱等可能有入口优势
- 商业智能场景: 数据和行业 know-how 是护城河

- OpenAI 的 1P-3P 产品法则

- 只要一两个人 (1P) 开发的产品就自己 (first Party) 做
 - 例: OpenAI API、ChatGPT、GPTs Store
- 需要三个人 (3P) 以上开发的产品就让第三方 (third Party) 做
 - 例: 行业模型、复杂任务规划求解、复杂的记忆系统

- 下面几个“有用 AI”的例子都是一两个人可以开发的

- 公司专注于基础模型能力和 infra, 相信 scaling law, 而不是不断打补丁

例 1: 帶着 AI Agent 一起出去玩

B Brian 02:25
Voice message 00:09

M maggie 02:25
🗣️: Yeah, I have a one day's visit, you have too much places to visit, could you shorten it to a one day's visit?

In reply to [this message](#) 02:25

Certainly! If you have just one day in Joshua Tree National Park, it's best to focus on a few key highlights that offer a variety of experiences. Here's a suggested itinerary for a one-day visit:

Morning:

- Start at the Visitor Center: Pick up a map, pay the entrance fee, and get the latest information on trail conditions and any potential closures.
- Hidden Valley: Begin with a short hike in Hidden Valley to get a feel for the iconic Joshua trees and rock formations. The loop trail here is approximately one mile and easy to manage.

Midday:

- Keys View: Drive up to Keys View for a panoramic vista of the Coachella Valley. This is a quick stop that doesn't require much walking, so it's perfect for a midday overview of the park's stunning geography.
- Cap Rock: On your way back from Keys View, make a short stop at Cap Rock, which is a quick and easy walk around some interesting rock formations.


B Brian



What's this sign?

M maggie
In reply to [this message](#)
No camping allowed.

B Brian 08:44



Do you know where is the photo taken?

M maggie 08:44
In reply to [this message](#)

The photo you've provided appears to be of the Stanford Memorial Church located on the campus of Stanford University in Stanford, California. The church is known for its distinctive architecture, beautiful mosaics, and it serves as a non-denominational place of worship. It's a landmark on the university campus and is frequently visited by students, faculty, and tourists.

例 2：企业 ERP 助手：从 GUI 到 LUI

只读查询

1. 平均每个员工在公司在职多久？
2. 公司每个部门有多少在职员工？
3. 哪个部门的员工平均级别最高？
4. 每个部门今年和去年各新入职了多少人？
5. 从 2020 年 3 月到 2021 年 5 月，A 部门的平均工资是多少？
6. 去年 A 部门和 B 部门的平均工资哪个高？
7. 今年每个级别的员工平均工资分别是多少？
8. 入职时间一年内、一年到两年、两年到三年的员工最近一个月的平均工资是多少？
9. 从去年到今年涨薪幅度最大的 10 位员工是谁？
10. 有没有出现过拖欠员工工资的情况，也就是某个月员工在职但是没有发薪？

增删改

1. 给小明升一级。
2. 新增一名员工，员工姓名是张三，技术部，3 级，入职日期是今天。
3. 删除名为小明的员工及其所有发薪记录。
4. 给 A 部门的员工发 2024 年 2 月的工资，工资跟上个月的相同。
5. 给所有 2024 年新入职的员工发入职奖金，奖金金额是员工级别乘以 10000。
6. 裁掉 A 部门的所有员工，也就是把离职日期设置成今天。
7. 给所有拖欠员工工资的情况补发工资，也就是如果某个月员工在职但没有发薪，就按照他拖欠工资前最后一次发薪的工资来补发。

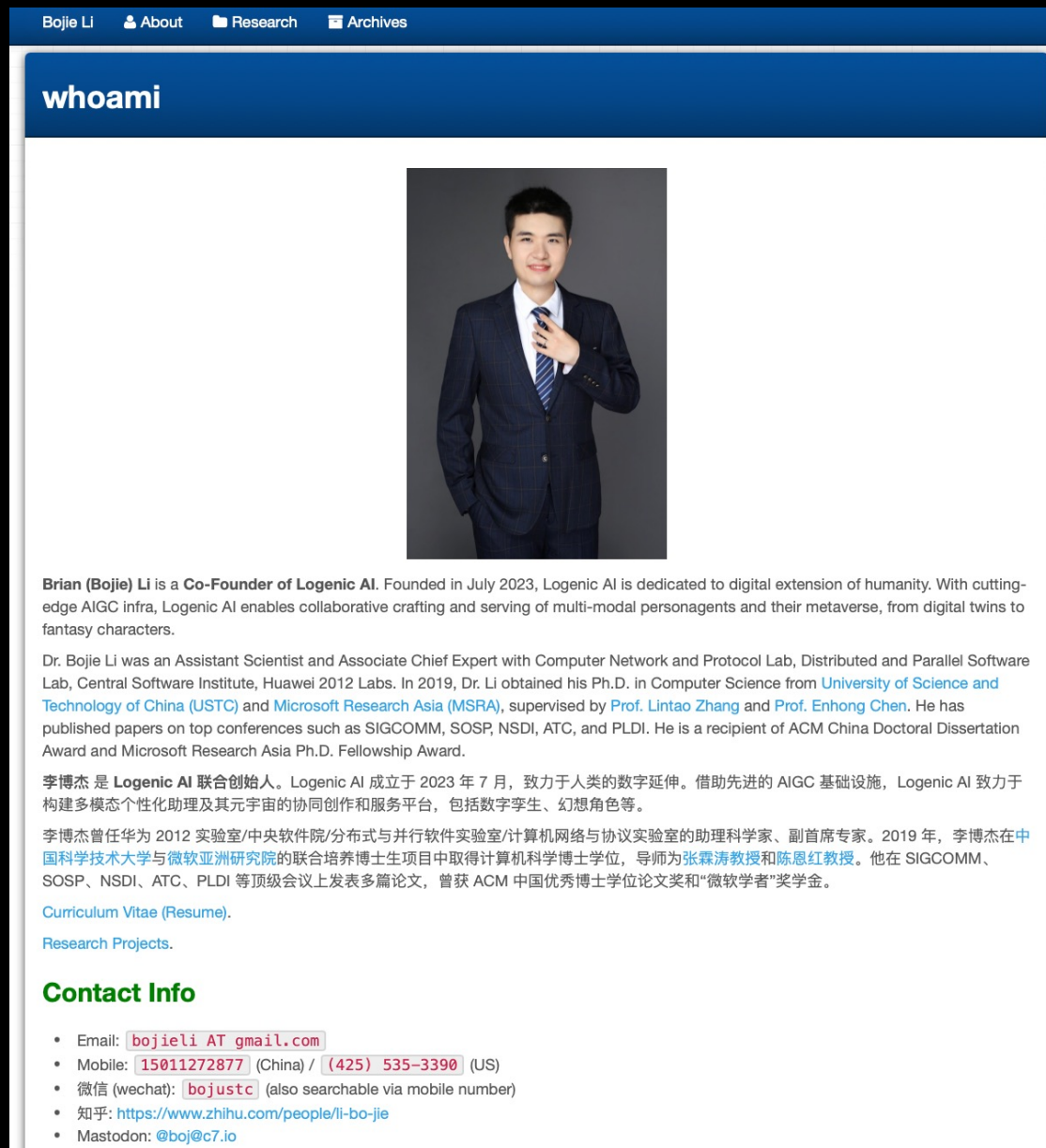
例 3：大模型采集数据

• 传统爬虫方法

- 每个院系主页/实验室主页/个人主页需要分别开发解析规则，每种类型的网页约需开发 1 小时
- 复杂的非结构化信息难以提取
 - 例如这种格式的邮件地址 bojieli AT gmail.com


• 大模型采集方法

- 通用性强，不需要任何定制开发
- 缺点：成本较高，GPT-4 读一个网页需 0.01 ~ 0.1 美金



Bojie Li About Research Archives

whoami



Brian (Bojie) Li is a **Co-Founder of Logenic AI**. Founded in July 2023, Logenic AI is dedicated to digital extension of humanity. With cutting-edge AIGC infra, Logenic AI enables collaborative crafting and serving of multi-modal personagents and their metaverse, from digital twins to fantasy characters.

Dr. Bojie Li was an Assistant Scientist and Associate Chief Expert with Computer Network and Protocol Lab, Distributed and Parallel Software Lab, Central Software Institute, Huawei 2012 Labs. In 2019, Dr. Li obtained his Ph.D. in Computer Science from [University of Science and Technology of China \(USTC\)](#) and [Microsoft Research Asia \(MSRA\)](#), supervised by [Prof. Lintao Zhang](#) and [Prof. Enhong Chen](#). He has published papers on top conferences such as SIGCOMM, SOSP, NSDI, ATC, and PLDI. He is a recipient of ACM China Doctoral Dissertation Award and Microsoft Research Asia Ph.D. Fellowship Award.

李博杰是 **Logenic AI** 联合创始人。Logenic AI 成立于 2023 年 7 月，致力于人类的数字延伸。借助先进的 AIGC 基础设施，Logenic AI 致力于构建多模态个性化助理及其元宇宙的协同创作和服务平台，包括数字孪生、幻想角色等。

李博杰曾任华为 2012 实验室/中央软件院/分布式与并行软件实验室/计算机网络与协议实验室的助理科学家、副首席专家。2019 年，李博杰在 [中国科学技术大学与微软亚洲研究院](#) 的联合培养博士生项目中取得计算机科学博士学位，导师为 [张霖涛教授](#) 和 [陈恩红教授](#)。他在 SIGCOMM、SOSP、NSDI、ATC、PLDI 等顶级会议上发表多篇论文，曾获 ACM 中国优秀博士学位论文奖和“微软学者”奖学金。

[Curriculum Vitae \(Resume\)](#).

[Research Projects](#).

Contact Info

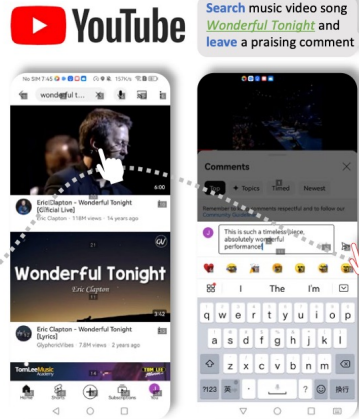
- Email: bojieli AT gmail.com
- Mobile: [15011272877](tel:15011272877) (China) / [\(425\) 535-3390](tel:(425) 535-3390) (US)
- 微信 (wechat): [bojustc](#) (also searchable via mobile number)
- 知乎: <https://www.zhihu.com/people/li-bo-jie>
- Mastodon: [@boj@c7.io](https://mstdn.io/@boj@c7.io)

例 4: RPA: 用 AI Agent 操作手机/电脑



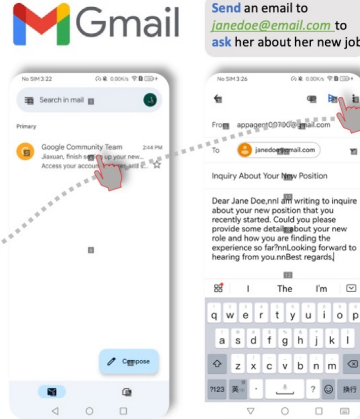
Chat with the user **Agent Liu** and answer his question

Observation: The screenshot shows the Telegram app, a conversation from "Agent Liu" with a message...
Thought: To complete this task, I should input the text with this answer into the input field labeled with the tag 'Z'.
Action: text("The Toronto Raptors won the 2019 NBA...")



Search music video song **Wonderful Tonight** and leave a praising comment

Observation: The screenshot shows a YouTube search result for "Wonderful Tonight"...
Thought: To complete this task, I should insert a praising comment into the text input field labeled 'Z'.
Action: text("This is such a timeless piece...")



Send an email to **janedoe@email.com** to ask her about her new job

Observation: The screenshot shows a smartphone's email application...
Thought: To complete this task, I should type the content of the email asking Jane Doe about her new job.
Action: text("Dear Jane Doe,\n\nI am writing to...")



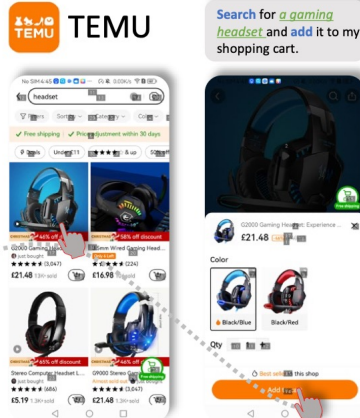
Improve the quality of this image until it looks good to you

Observation: The screenshot shows a photo editing app on a smartphone with an image displayed on the screen...
Thought: To complete this task, I should adjust the noise reduction level using the slider.
Action: Action: swipe(23, "right", "medium")



Set an alarm at **12:30 pm** every Friday and Sunday, and disable the vibration

Observation: The image shows a screenshot of an alarm application on a smartphone...
Thought: To complete this task, I should disable the vibration feature by tapping on the UI element.
Action: tap(17)



Search for a gaming headset and add it to my shopping cart.


Observation: The screenshot displays a shopping app with a variety of headsets listed...
Thought: To complete this task, I should add the selected headset into my shopping cart.
Action: tap(14)

RPA = Robotics Process Automation, 机器人流程自动化

AppAgent: Multimodal Agents as Smartphone Users, Tencent

例 4: RPA 方案: 视觉 VS 元素树

视觉方案



Observation

The screenshot shows a map application with the location "Tencent Shanghai Branch" pinpointed with a red marker. There are two buttons at the bottom of the screen, one labeled "Directions" and the other labeled "Start", with numeric tags 1 and 3 respectively.

Thought

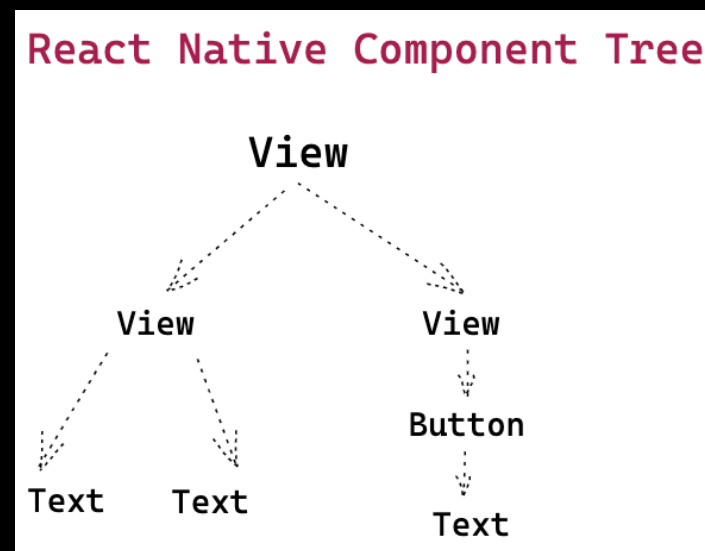
To complete the given task, which is to navigate to Tencent Shanghai Branch, I should tap the "Directions" button to initiate the route finding process.

Action

tap(2)

- 优点: 仅依赖屏幕截图, 通用性强
- 缺点: 细小屏幕组件识别不准确; 大块文字识别需 OCR 辅助; 成本较高, 特别是对于需要滚动才能显示完整的界面

元素树方案



- 优点: 识别准确率高; 成本较低; 无需 OCR 和视觉大模型
- 缺点: 需操作系统底层 API 权限; 大模型缺少元素树理解能力, 需要数据做预训练或微调; 元素树往往较大, 需要筛选可视部分输入到大模型

例 5：会议/生活记录器

- 会议中刚溜了一会儿号，就被老板 cue 到，一脸懵
 - 实时会议内容转写和总结（Zoom 和腾讯会议已经支持）
 - 关注话题提醒
- 今年都去过哪些地方？
 - 虽然各类 App 都有记录，但数据是烟囱化的，无法导出，无法聚合各类 App 的数据来做分析
 - AI Agent 可以通过 RPA 或 Intent-based API 方式收集生活记录
- Rewind.AI 的录屏和录音吊坠是我很喜欢的产品
 - 其实大厂信息安全部门一直在用类似的方法检测泄密
 - 隐私是记录生活最大的顾虑，本地化算力/隐私计算是必由之路

慢思考 (1): 思维链回答数学题

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

AI 回答数学题需要时间思考：绿色部分，“慢思考”，属于中间状态而非输出

慢思考 (2): 多步网络搜索回答难题

How many storeys are in the castle David Gregory inherited?

Vanilla LM

LM: Castle Gregory has three storeys.

✗ Hallucinates
a fictitious castle

Retrieve-
then-Read

RM: "St. Gregory Hotel is a nine-floor boutique hotel in D.C..."

LM: St. Gregory Hotel has nine storeys.

✗ Retrieves a
different building

Multi-Hop
DSP Program

LM: "Which castle did David Gregory inherit?"

RM: "David Gregory inherited Kinnairdy Castle in 1664..."

LM: "How many storeys does Kinnairdy Castle have?"

RM: "Kinnairdy Castle is a tower house, having five storeys..."

LM: Kinnairdy Castle has five storeys.



Multi-hop QA: 单次搜索不能得到答案时, 需要将问题拆分为多个子阶段

慢思考 (2): 复杂任务规划分解





- 论文第二章相比相关工作 X 的贡献是什么?
 - 如何找到第 2 章的所有内容?
 - 如何找到相关工作 X, 并总结它的贡献?
- 查询洛杉矶今天的天气
 - 简单解析 HTML 或文本难以区分网页中不同日期和不同地区的天气
 - 视觉理解是终极方案

Brea, CA As of 5:05 pm PDT


80°
Sunny
Day 84° • Night 61°

Watch: Scotland Turned Into 'Winter Wonderland' But It's Not Snow


Today's Forecast for Brea, CA

Morning	Afternoon	Evening	Overnight
76°  --	82°  --	66°  2%	62°  12%

Next Hours

Advertisement: 5X points on eligible prepaid travel on AmexTravel.com. 

Weather Today in Brea, CA

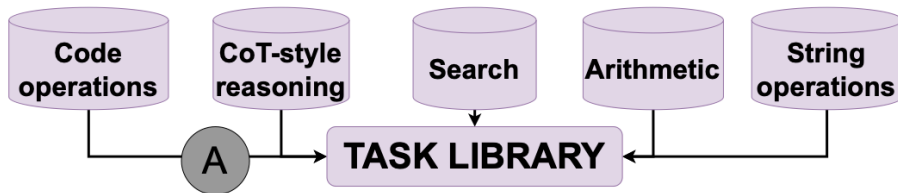
Feels Like **81°** 
7:01 am 6:10 pm

High / Low	--/61°	Wind	8 mph
Humidity	44%	Dew Point	56°

慢思考 (3): 自动调用多种工具分步求解

New Task (Physics QA) Answer this high-school physics question

Input: Hector yanks on the chain with a 72.0 N force at an angle of 35.0° above the horizontal. Determine the horizontal components of the tension force.



B TOOL LIBRARY



Solve these arithmetic problems using python code
Input: Viola had 167 breads. Nancy took 137 from him. How many does Viola have now?
Q1: [generate code] Write down arithmetic as python code
#1: viola_bought = 167, nancy_took = 137
ans = viola_bought - nancy_took
Q2: [code execute] Execute snippet **#2:** 30 **Q3:** [EOQ] **Ans:** No

Does the sentence contain an anachronism? Yes/No.
Input: President George H. W. Bush called his generals at the outset of the Gulf War.
Q1: [search] When was President George H. W. Bush, president?
#1: George H. W. Bush's tenure started on January 20, 1989, and ended on January 20, 1993.
Q2: [search] When was the Gulf War fought? **#2:** The Gulf War was a 1990–1991
Q3: [subquestion] Could these entities have co-existed? **#3:** Yes. Their time periods intersect.
Q4: [generate output] Is this an anachronism? **#4:** No **Q5:** [EOQ] **Ans:** No

Input: Hector yanks on the chain with a 72.0 N force at an angle of 35.0° above the horizontal. Determine the horizontal components of the tension force.
Q1: [search] What is the formula for the horizontal component of the tension force?
#1: The formula for the horizontal component of the tension force is $T\cos\theta$. The horizontal component (F_x) can be calculated as $F_{tens}\cosine(\theta)$ where θ is the angle which the force makes with the horizontal in radians.

LLM

Input: ... **Q1:** [search] ...
#1: ... can be calculated as $F_{tens}\cosine(\theta)$ where θ is ...
Q2: [generate code] Use the formula $F_x = F_{tens}\cosine(\theta)$ to solve: Hank ...
#2: $T = 72.0$, $\theta = 35.0$
 $radians = \text{math.pi} * \theta / 180$
 $F_x = T * \text{math.cos}(radians)$

OpenAI Codex

Input: ... **Q1:** [search] ... **#1:** ...
Q2: [generate code] Use the formula $F_x = F_{tens}\cosine(\theta)$ to solve: Hank ...
#2: ... $F_x = T * \text{math.cos}(radians)$
Q3: [code execute] Execute the python code and get the value of "Fx"
#3: 58.9789
Q4: [EOQ] **Ans:** 58.9789



告诉 AI 调用工具的参考例子

依次调用 Google, OpenAI Codex 和 Python 解决问题

慢思考 (3): 让 AI 学会使用上万种工具

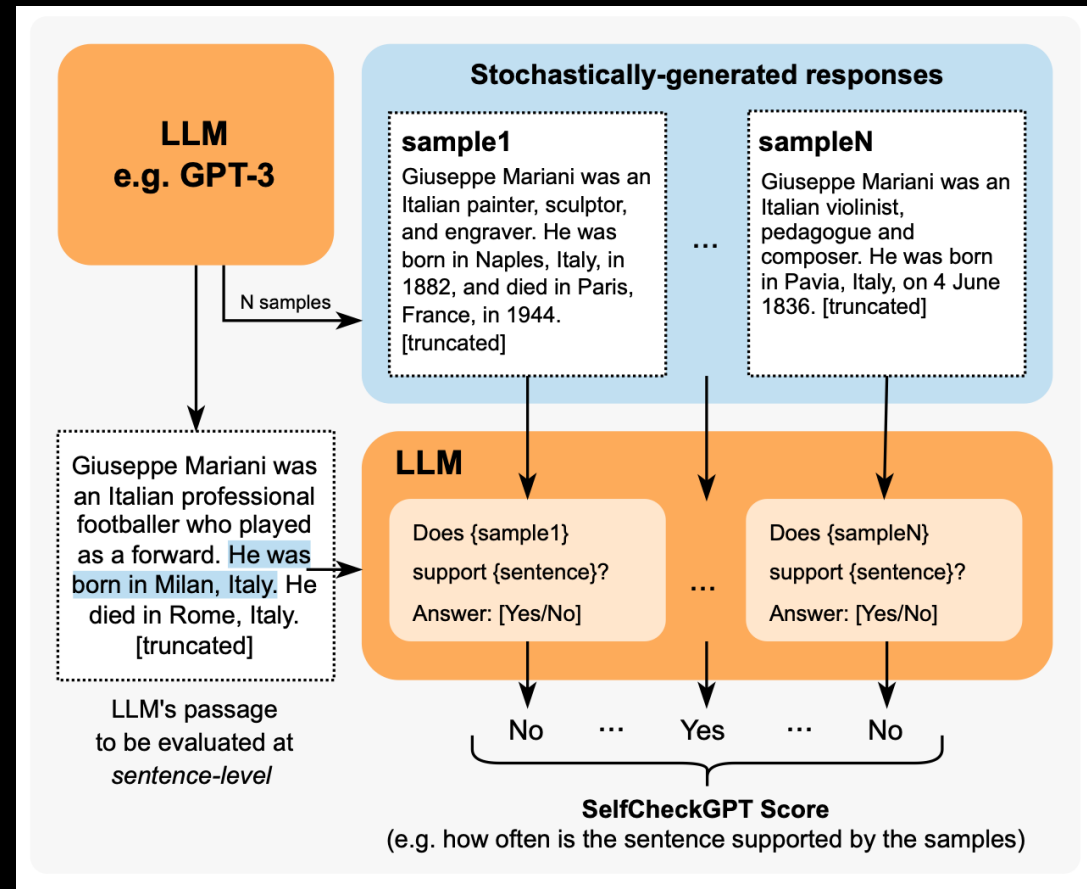
- GPT Store 路线（工具调用大模型） vs ChatGPT 路线（大模型调用工具）
 - GPT Store 路线是用户指定使用何种工具
 - ChatGPT 路线相当于是把工具的说明书都摊开在桌面上，用时再查找
- 大模型如何学会自动使用 GPT Store 中的上万种工具？
- 观点1：工具使用属于过程记忆，不是语言可以明确描述的
 - 使用 Fine-tuning 方法，甚至在预训练时加入工具使用的语料
- 观点2：工具使用可以用代码形式表达，因此属于代码生成能力
 - 使用 RAG 方法获取到工具使用的代码

减少幻觉的工程方法

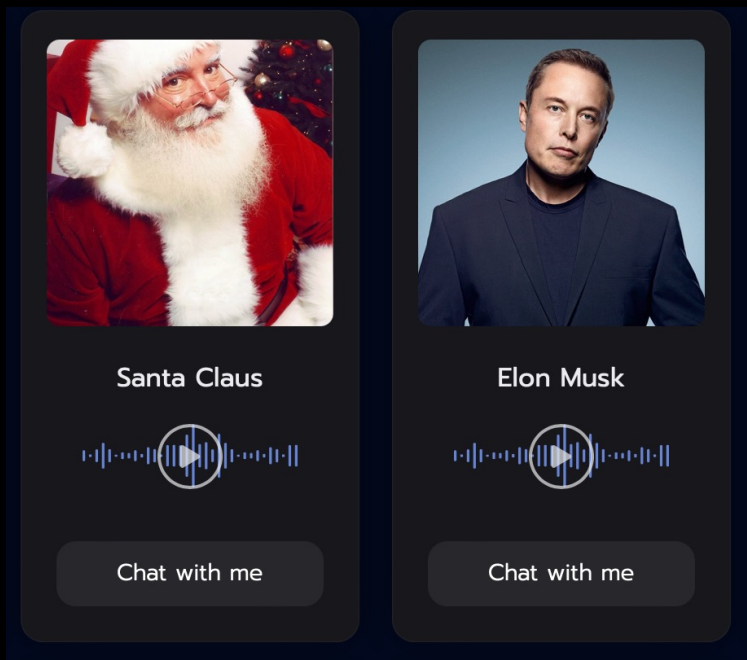
事实性校验 (Factual Checking)



多次生成 (SelfCheckGPT)



有用的 AI 价值更高



语音闲聊

收入：1 元/小时
用 GPT-3.5 成本：5 元/小时
用 Mixtral 8*7B 成本：1 元/小时
大多数现有平台都在赔钱



在线教育

收入：10 元/小时
用 GPT-3.5 成本：5 元/小时
用 Mixtral 8*7B 成本：1 元/小时
相对容易赚钱



心理咨询、法律咨询

收入：100 元/小时
用 GPT-4 成本：50 元/小时
用 GPT-3.5 成本：5 元/小时
质量和品牌是关键

让大多数人用得起 AI

- Model Router

- 简单问题用小模型（如 Mistral 7B），复杂问题用大模型（如 Mixtral 8*7B, LLaMA-2 70B 甚至 GPT-4）
- 小模型如何知道自己回答不了，把问题交给大模型，而不是输出幻觉？

- Inference Infra

- 大量 LoRA 批量推理
- FlashAttention 等优化
- 持久化 KV Cache

- Datacenter Infra

- 自建算力平台，使用消费级显卡做推理

成本：自建 Infra VS 调用 API

GPT-4 API

8K input tokens * \$0.015 / 1K input tokens
+ 0.5K output tokens * \$0.03 / 1K output tokens
= **\$135** / 1K queries

GPT-3.5 API

(8K input tokens * \$0.001 / 1K input tokens
+ 0.5K output tokens * \$0.002 / 1K output tokens)
= **\$9** / 1K queries

Mixtral 8*7B MoE API (Together.AI)

(8K input tokens + 0.5K output tokens)
* \$0.0006 / 1K tokens
= **\$5.1** / 1K queries

Mixtral 8*7B MoE (self-hosted on 4*H100)

(8K input tokens / 10K prefill tokens/s
+ 0.5K output tokens / 50 output tokens/s / 64 parallel)
* \$2 / hour / GPU * 4 GPUs / 3600 s/hour
= **\$2.1** / 1K queries

Mistral 7B API (Together.AI)

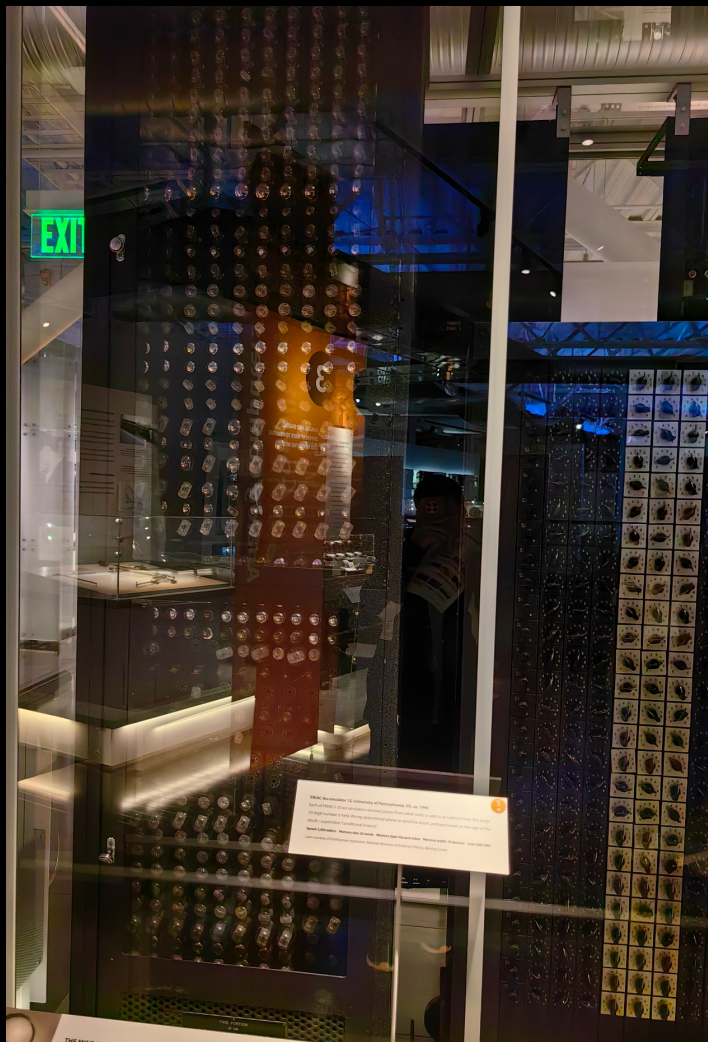
(8K input tokens + 0.5K output tokens)
* \$0.0002 / 1K tokens
= **\$1.7** / 1K queries

Mistral 7B (self-hosted on 4090)

(8K input tokens / 4K prefill tokens/s
+ 0.5K output tokens / 20 output tokens/s / 32 parallel)
* \$0.5 / hour / GPU / 3600 s/hour
= **\$0.39** / 1K queries

* API 价格计算于 2023 年 12 月，可能随时变动

大模型的成本一定会快速降低



计算机历史博物馆：ENIAC 每秒 5000 次加法，只有 20 words 的内存

最初版的 ChatGPT：每秒输出 30 个 token，4096 tokens 的内存

今天的 ChatGPT：每秒输出 150 个 token，16384 tokens 的内存

未来会不会有一天，可以每秒输出上万个 token，有上亿个 token 的内存？

最早的 LLaMA 65B：成本 10 美元 / 百万 token (TensorRT + A100)

今天的 Mixtral 8*7B：成本 0.3 美元 / 百万 token (vLLM + 4090)

未来会不会有一天，同等能力的模型可以跑在手机上？

超级智能：有效加速 VS 超级对齐

- 有效加速 (Effective Acceleration, e/acc)
 - AI 技术是中性的，关键在于使用它的人，不需要限制 AI 技术的发展
 - 人类肉体有很多物理上的限制，硅基可能是比碳基更合适的生命形式
 - 不应该用人类的价值观约束超级智能
 - Elon Musk: Humanity Is a Kind of 'Biological Boot Loader' for AI
- 超级对齐 (Superalignment)
 - AI 就像原子弹，需要成立类似国际原子能组织的机构，控制 AI 的发展以免威胁人类
 - 如何保证比人类更强大的 AI 始终遵循人类的意图，听从人类的指挥
 - Evaluation is easier than generation.
 - Jan Leike, Head of OpenAI Superalignment Team

开源模型还是闭源模型

- 最好的模型一定是闭源模型
- 但很多场景下开源模型就足够了
 - 例如泛娱乐场景的 Agent
- 没有基础模型能力的公司基于开源模型更容易建立护城河
 - 微调
 - 推理优化
 - 多模态
 - 基于 Embedding 的记忆
 - 流式推理（例：多个 Agent 语音聊天）
 - 本地化部署、终端设备、机器人
- 用户真正具有所有权的 AI Agent 一定是基于开源模型
 - 如果未来真的有数字生命，这对人类的命运是很关键的

数字生命：赋予每个人无限时间

- 人类社会的稀缺性来自物质或时间的稀缺性

- 很多名人很想跟粉丝一对一交流，但没有时间

- “The biggest consideration I take into doing content deals is how much time, it is going to take Stephen away from basketball and his family.”

-- Jeff Austin, on his process with his client Stephen Curry

- 人的生命只有一次，这就是为什么游戏如此流行

- Galgame, 乙女游戏, 体验多条时间线的多种可能

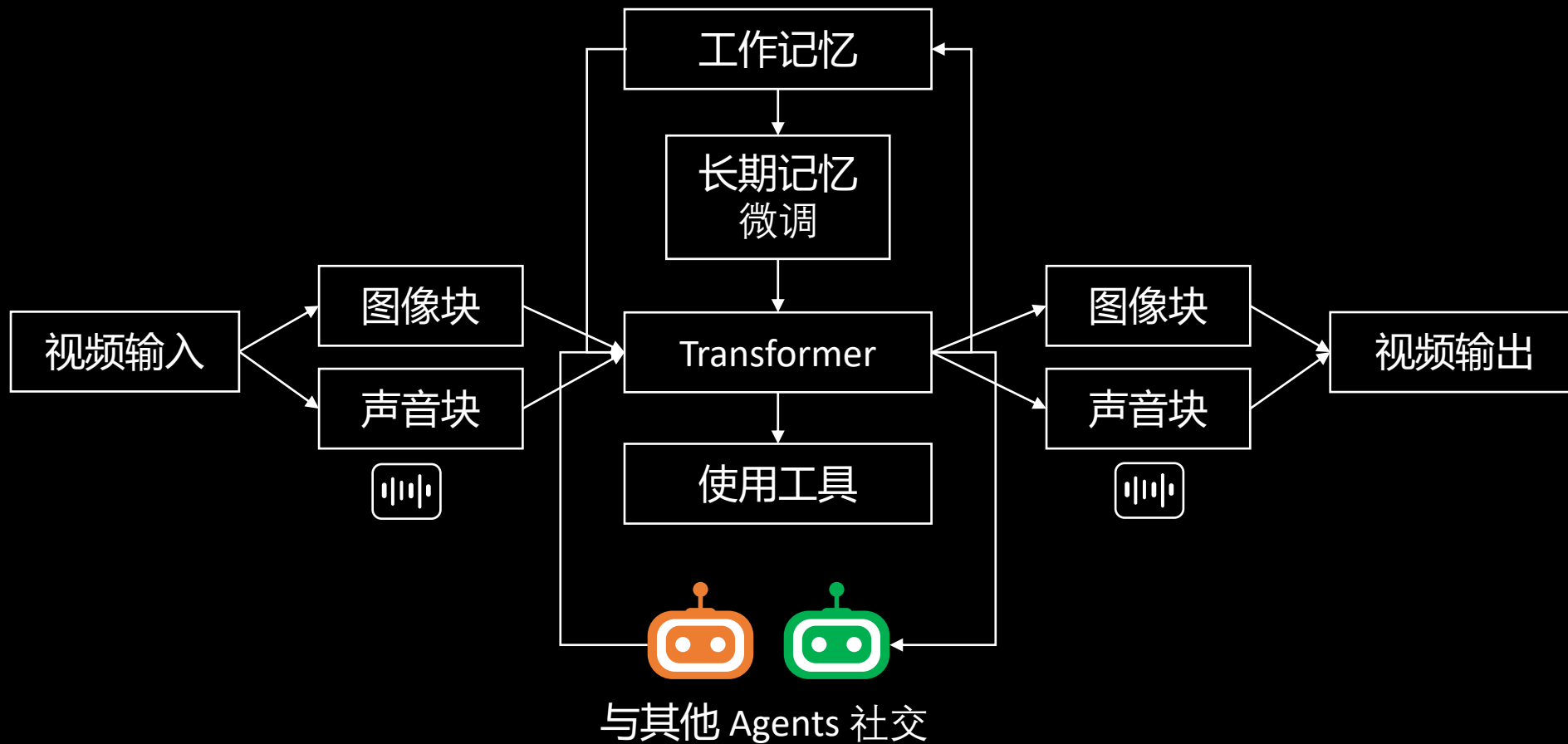
- 把生命变成无限的真的很好吗？

- AI is not a creature but a tool. – Sam Altman

- Digital extension of humanity – Our mission



我们距离数字生命还有多远？



去中心化是 AI Agent 的未来

- 如今的 AI Agent 数据属于中心化平台（OpenAI、Character AI）
 - 用户没有所有权，平台可以随时关停和篡改 AI Agent
 - 如果数字生命被一家公司掌握，这将是多么可怕的一件事
 - 用户没有经济动机，赚的所有钱都归平台，缺少利润分享机制
 - “用爱发电”导致 Character AI 平台上缺少高质量的 Chatbot，用户留存和付费意愿都很差，平台反而赚不到钱
- 去中心化（Decentralization）是 AI Agent 的未来
 - 运行在去中心化的算力和模型上，不用担心平台跑路
 - 每个用户完全拥有自己的 AI Agent 或者数字分身，保证隐私和所有权
 - 未来，有自主意识的数字生命也需要有独立的人权，不能被中心化控制
 - AI Agent 的创作者可以盈利，只需支付透明的去中心化算力和模型成本

赋予每个人无限时间
有趣的灵魂终会相遇