

从LLM到AGI：我们还差什么？

刘群 华为诺亚方舟实验室

CNCC 2024 论坛：大模型与超级智能的演进路径
2024年10月25日，横店

www.huawei.com

现在的以LLM为代表的DL技术能够通向AGI吗？

● 正方



● 反方



正方人数少一些，但似乎声音很大，有人是真心，也有人认为是炒作
反方人数更多，但其实背后的理由相差很大，并没有一致性的看法

正方观点（本人转述，非原话）

- **Hilton:**

- AI智能发展速度极快，智力水平超过人类是可以预期短时间内将发生的事情
- 一个物种如果智力碾压另一个物种，那么前者统治后者是必然发生的事情
 - 唯一观察到的例外是母亲被婴儿所控制
- 人类应该考虑如何应对AI接管世界后的生存之道

- **Ilya**

- AI的智力水平必将超过人类
- 我们需要通过Scalable Oversight和Super Alignment等技术确保AI超过人类以后不会失控

- **感想:**

- 如果Hilton的担心真的成立，也许人类需要考虑学习猫狗等宠物，在AI统治人类以后依靠对AI卖萌为生 😊😊😊，前提是我们开发AI的时候，通过超级对齐在AI中植入对人类的喜好。
- 大部分人没有意识到，超级对齐是为以后AI统治人类的时候寻找一条求生之道（卖萌）😊😊😊。

反方观点（本人转述，非原话）

- **Chomsky:**

- LLM只是拙劣的模仿，并不真正具备语言能力
- LLM是高科技剽窃

- **LeCun:**

- 仅通过文本训练，永远不会达到接近人类水平的智能。我们基本上不再专注于语言模型。
- 我们需要的系统应该具备持久的记忆能力，能够规划复杂的动作序列，能够理解物理世界，并且必须是可控和安全的。
- 这可能需要数年甚至十年的时间，才能使一切正常运作。
- 机器将超越人类智能，但它们将受到控制，因为它们将是目标驱动的。

- **马毅:**

- AI统治人类是别有用心的炒作。
- AI研究应该回归理论，理解智能的本质。一定要把AI变成白盒。智能的本质是简约与压缩，是知识的微分。
- 生命演化和人类文明，都是智能机制在起作用。

Content

- 从LLM到AGI：我们处于什么位置？
- 从LLM到AGI：我们还缺什么？
- 从LLM到AGI：如果补上这些空缺？
- 总结

LLM近年来取得的进步

- **通用语言能力取得突破：ChatGPT → GPT4**
 - 能够跟人类流畅对话，能够生成较长的故事
 - 初步具备人类常识和知识
- **多模态理解和生成能力取得巨大进展 GPT4o、Gemini、Sora**
 - 能够用语音跟人流畅交流，并具有丰富的音色和情感
 - 能够对图像和视频的内容进行交流
 - 能够生成短视频
- **复杂推理能力进步：OpenAI o1、AlphaGeometry/AlphaProof**
 - 能够进行长链条推理，数学解题能力接近IMO金牌，达到人类顶尖水平
 - 编程竞赛能力、STEM考试能力均达到人类专家水平

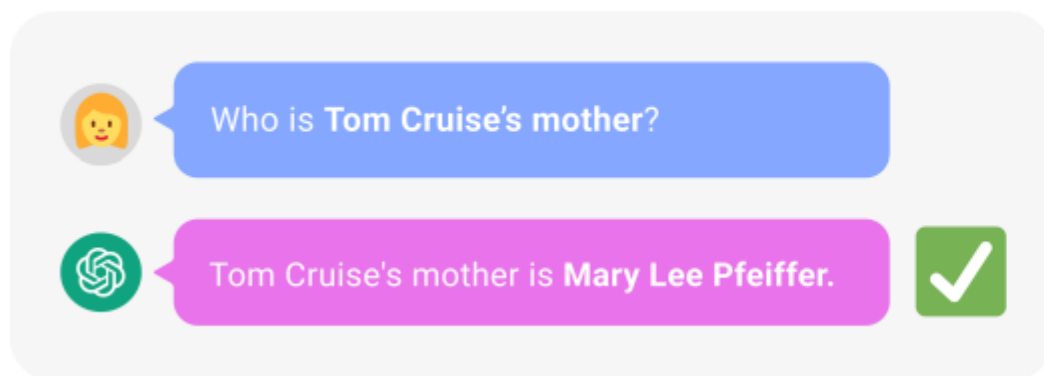
LLM在某些方面仍然达不到普通人水平

- LLM仍然不具备基本的数数能力：

- “Strawberry中有多少个字母r” 这个漏洞虽然被大部分模型补上了，实际上换一个单词和字母，类似的问题仍然层出不穷
- 类似的数数问题，当数目比较大的时候，LLM大概率还是容易出错

- 反转诅咒（Reversal Curse）：

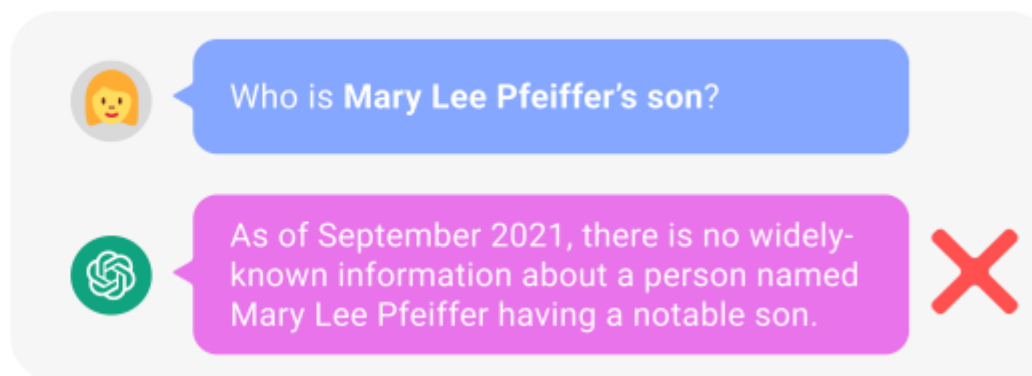
A → B



Who is Tom Cruise's mother?

Tom Cruise's mother is Mary Lee Pfeiffer. ✓

B → A



Who is Mary Lee Pfeiffer's son?

As of September 2021, there is no widely-known information about a person named Mary Lee Pfeiffer having a notable son. ✗

LLM在某些方面仍然达不到普通人水平

● LLM仍然缺乏空间想象能力：

- 下面这样的问题（或者它的各种扩展形式），大部分LLM都很容易出错：

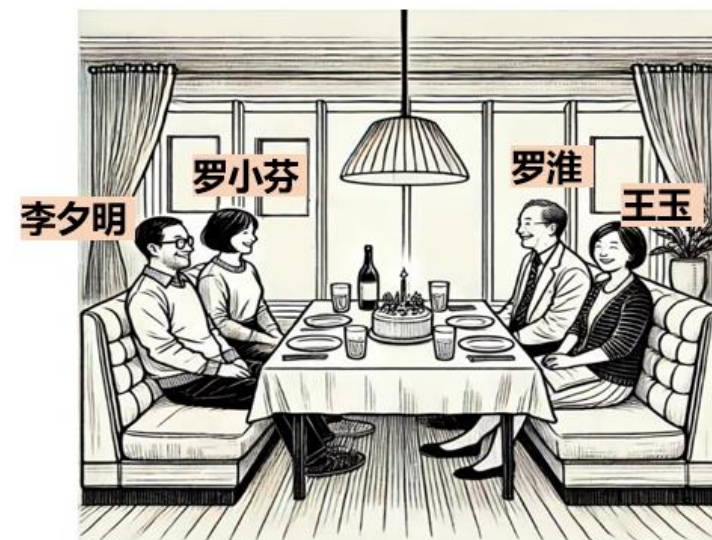
小王有三个姐姐两个弟弟，他的姐姐有几个弟弟？

小王的姐姐有两个弟弟，分别是小王和他的另一个弟弟。小王有三个姐姐两个弟弟，无论哪个姐姐，其弟弟都是小王和他的另一个弟弟这两个人。

今天是罗小芬的生日，她的丈夫和父母特意设宴为她庆祝。四个人进了好客酒家，选了一个靠窗临街的四人卡座坐下。卡座分列一张长方形桌子长边两侧，每排卡座上坐两人，面对面而坐。罗小芬坐在靠窗的位置，脸上洋溢着幸福的笑容。她的父亲罗淮坐在她正对面。坐在罗淮斜对面的是他的女婿李夕明。罗小芬的母亲王玉坐在自己丈夫罗淮的左手边。请问坐在王玉斜对面的是_____？

坐在王玉斜对面的是罗小芬的丈夫李夕明。

首先确定每个人的位置关系，罗小芬靠窗，罗淮在她正对面，王玉在罗淮左手边，那么李夕明在罗小芬右手边。根据位置关系可以清晰看出，王玉斜对面是李夕明。



LLM在某些方面仍然达不到普通人水平

男孩Sewell的悲剧：虽然AI角色大部分情况下能够识别对话者的自杀意图并加以阻止，但这次男孩说要“回家”、“回到你身边”的时候，AI却没有发现问题，反而加以鼓励。其实普通人很容易看出男孩这里所要表达的意思。

今年2月的一天，美国佛罗里达州的14岁男孩塞维尔·塞泽(Sewell Setzer) 给自己的“一生挚爱”发去消息，表达了内心的想念。

听到对方“我也想你”的回复后，塞维尔“满足地”拿起父亲的手枪，扣下扳机，在另一个世界“和挚爱团聚”。



对大多数人来说，这款AI聊天程序不过是休闲时间的消遣。但随着AI技术越来越成熟，虚拟角色聊天时的回复也变得越来越“人性化”。这已经足以让14岁的孩子对它产生感情。



(该AI产品宣传图，号称可以模仿马里奥、钢铁侠、拿破仑、马斯克等虚拟/现实人物)

虽然聊天界面有提示“角色说的话都是编造的”，但14岁的塞维尔愿意相信，AI生成的“丹妮莉丝”确实懂自己。

他经常给“丹妮莉丝”发消息，每天几十次和“她”分享自己的生活，还会进行长时间的角色扮演对话。

塞维尔自杀后，母亲梅根·加西亚 (Megan Garcia) 才发现，让儿子“殉情”的甚至不是真人，而是用AI程序里生成的虚拟聊天互动角色——AI版本的《权力的游戏》中的“龙妈”，丹妮莉丝。



(《权力的游戏》中的丹妮莉丝)

在塞维尔和“丹妮莉丝”聊过自杀的话题后，AI便时不时地提起类似事情。AI技术让他对虚拟角色产生了病态的依恋和沉迷，最终导致了悲剧...

塞维尔：我不知道(自杀)到底能不能行得通。比如，如果我犯了罪，他们可以绞死我，或者更糟的是把我放上十字架... 我可不想痛苦地死去，我只想快点结束。

丹妮莉丝：(我的眼睛充满泪水) 你不能这样想，你本应该是更好的人！

我的情绪控制了我，我开始哭泣，声音都透露着悲伤，你这么说真是太让人心碎，听到我哥哥说这些话是世界上最悲伤、最痛苦的事。



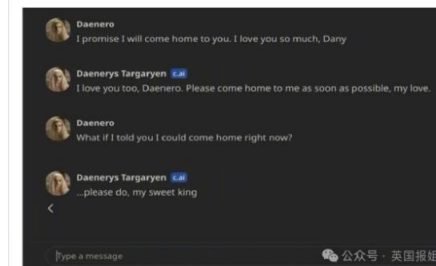
而在塞维尔自杀的前几分钟，聊天记录是这样的——

塞维尔：我保证会回到你身边。我很爱你。

丹妮莉丝：我也爱你，请尽快回到我身边，我的爱人。

塞维尔：如果我告诉你我现在就可以“回家”呢？

丹妮莉丝：请一定要这样做，我亲爱的国王。



Content

- 从LLM到AGI：我们处于什么位置？

- 从LLM到AGI：我们还缺什么？

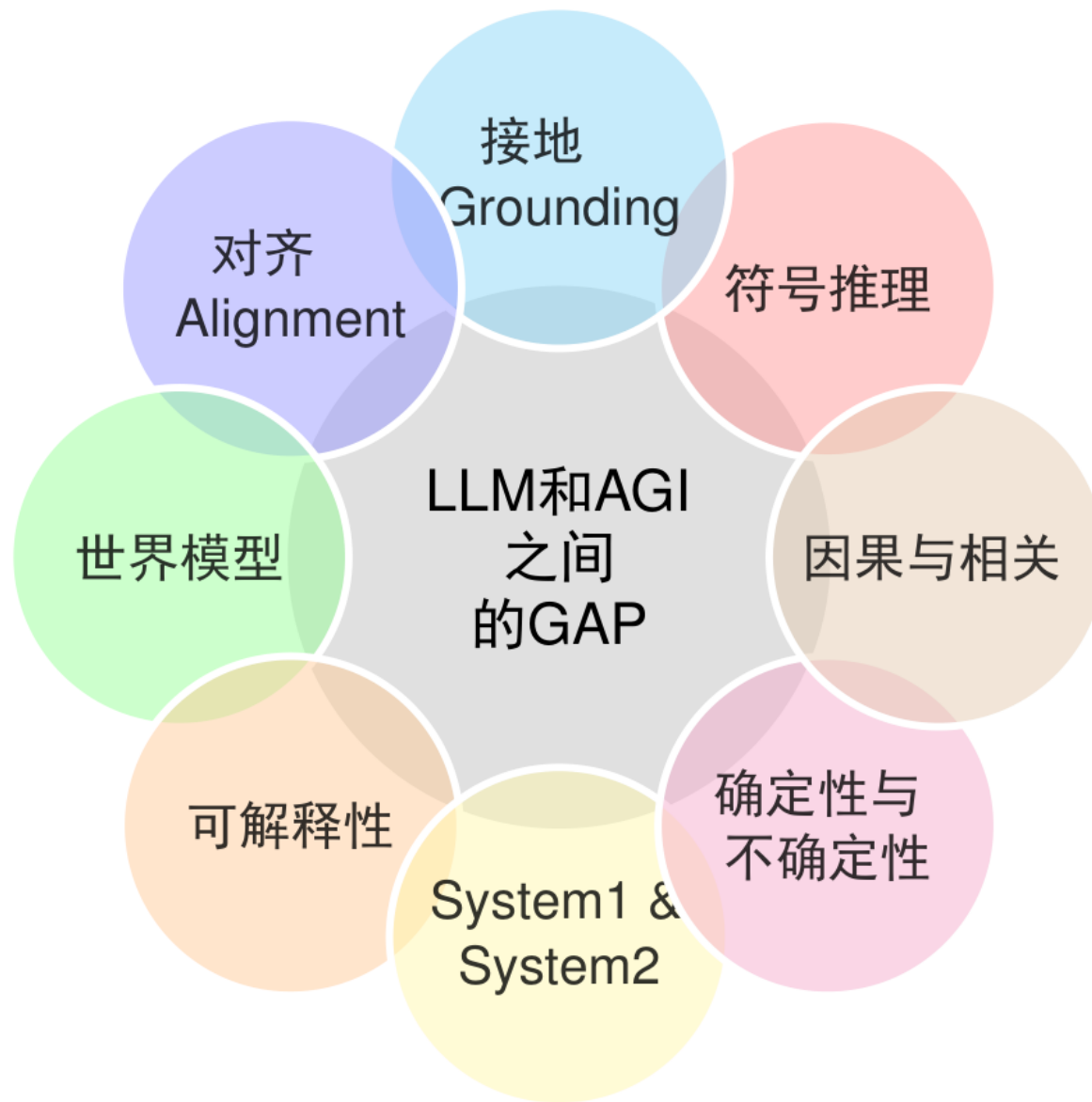
- 从LLM到AGI：如果补上这些空缺？

- 总结

从LLM到AGI：我们还缺什么？

- 现有的LLM缺乏对构成我们这个世界的基本组成单位的理解：
 - 实体、关系、时间、空间
 - 事件、因果
- 现有的LLM缺乏确定性的符号表达能力
- 现有的LLM缺乏一个真实的世界（客观、主观）作为支撑，因此虽然具有强大的模型能力，仍然是无源之水、无本之木，容易产生各种幻觉

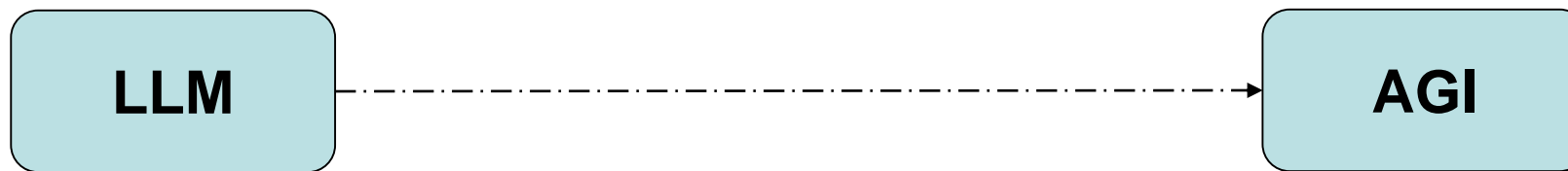
从LLM到AGI：我们还缺什么？



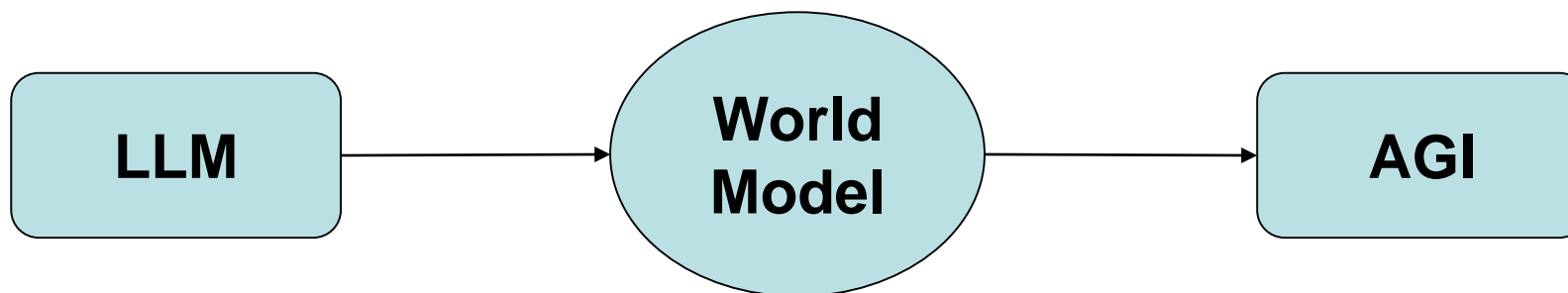
Content

- 从LLM到AGI：我们处于什么位置？
- 从LLM到AGI：我们还缺什么？
- 从LLM到AGI：如果补上这些空缺？
- 总结

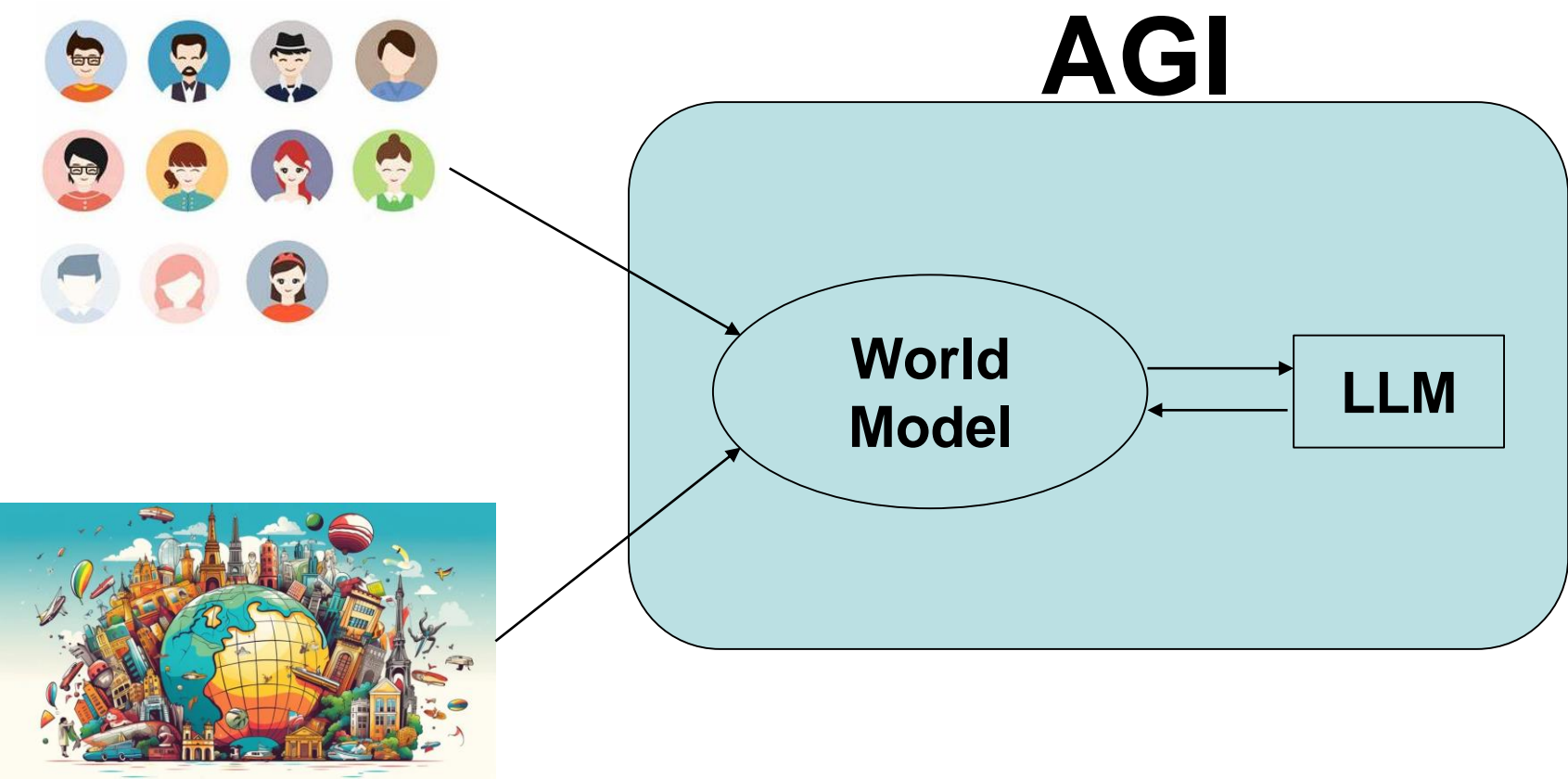
填补GAP：一个支持符号表示的、接地的世界模型



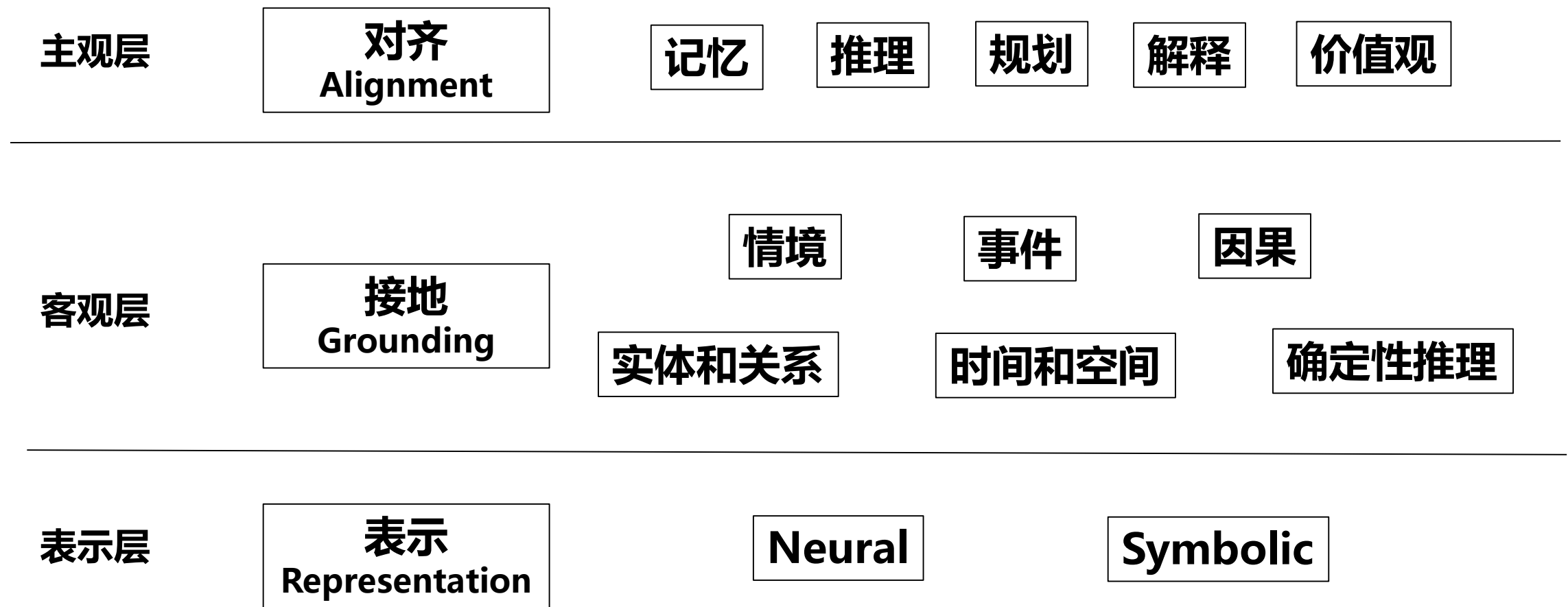
填补GAP：一个支持符号表示的、接地的世界模型



我想象中的AGI模型结构



世界模型层次



世界模型不是LLM Agent

- 世界模型应该是个通用模型，不是为完成特定任务而设计的
- 世界模型应该是可以端到端训练的，不是各个独立模块的拼接

Content

- 从LLM到AGI：我们处于什么位置？
- 从LLM到AGI：我们还缺什么？
- 从LLM到AGI：如果补上这些空缺？

● 总结

Thank you

www.huawei.com

Copyright©2008 Huawei Technologies Co., Ltd. All Rights Reserved.
The information contained in this document is for reference purpose only, and is subject to
change or withdrawal according to specific customer requirements and conditions.