稀土掘金

2023 JUEJIN 2023 TECH TALK
掘金年度技术演讲

大语言模型技术现状与发展趋势的思考
Thoughts on LLM Technology Trends

刘群 | 华为诺亚方舟实验室语音语义首席科学家

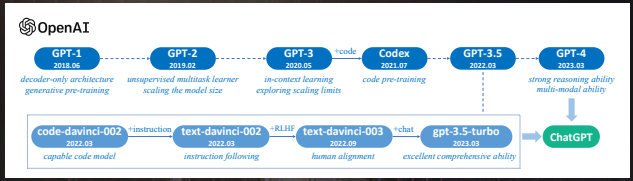
NOAH'S ARK LAB

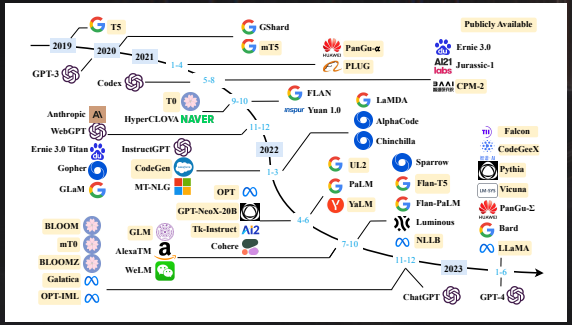
HUAWEI

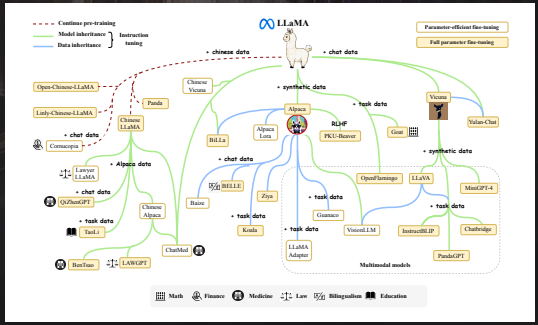
Content

业界大模型概览

- 2023年是大模型爆发的一年
- 2023年又被称为AGI元年
- 大模型已经深刻影响了AI
- 大模型还将深刻影响我们的社会

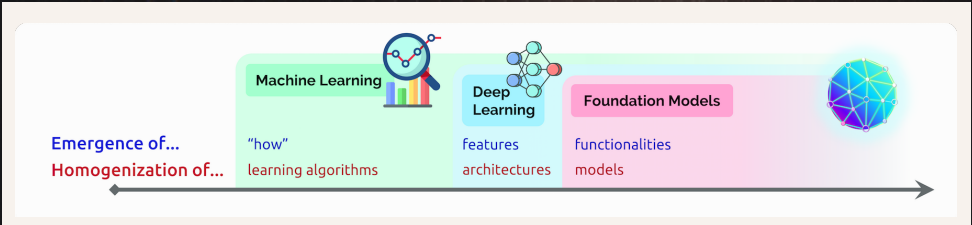






Zhao, et al. “A Survey of Large Language Models.” arXiv2303.18223

基础模型的涌现和同质化 Emergence and homogenization



Bommasani et al., On the Opportunities and Risks of Foundation Models, arXiv:2108.07258 [cs.LG]

	Machine Learning	Deep Learning	Foundation Models
Emergence	how a task is performed	the high-level features used for prediction	advanced functionalities such as in-context learning
Homogenization	learning algorithms (e.g., logistic regression)	model architectures (e.g., Convolutional NNs)	the model itself (e.g., GPT-3)

语言模型的技术同质化：预训练语言模型 vs. 大语言模型

	预训练语言模型 Pre-trained Language Models (PLMs)	大语言模型 Large Language Models (LLMs)
典型模型	ELMo, BERT, GPT	GPT-2, GPT-3
模型架构	BiLSTM, Transformer	Transformer
	Encoder, Encoder-decoder, Decoder	Decoder
注意力机制	Bidirectional、Unidirectional	Unidirectional
训练方式	Mask & Predict Autoregressive Generation	Autoregressive Generation
擅长任务类型	NLU	NLU & NLG
模型规模	0.1-1B parameters	1Billion-xTrillion parameters
下游任务应用方式	Fine-tuning	Prompting & Fine-tuning & RLHF
涌现能力	Inductive Transfer Learning	Zero-shot Learning Few-shot/In-context Learning Chain-of-Thought

语言模型的能力涌现：The scale matters

上下文学习（零样本/少样本学习）

Zero-shot

The model predicts the answer given only a natural language discription of the task. No gradient updates are performed.

```
Translate English to French:        ← task description
cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

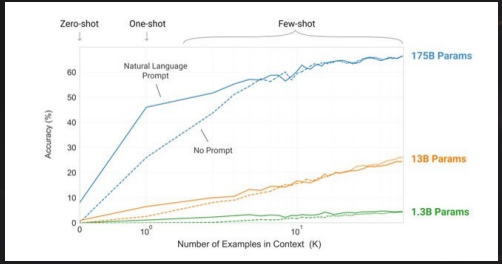
```
Translate English to French:        ← task description
sea otter => loutre de mer          ← example
cheese =>                           ← prompt
```

Few-shot

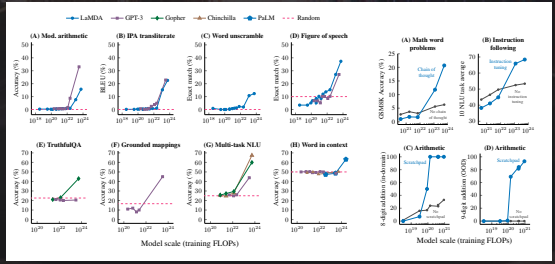
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:        ← task description
sea otter => loutre de mer          ← examples
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>                           ← prompt
```

上下文学习的能力涌现



其他能力涌现



大语言模型训练的尺度定律（Scaling Law）

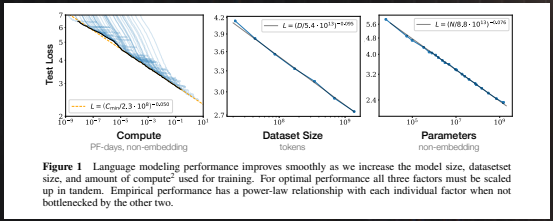


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Kaplan et al. “Scaling Laws for Neural Language Models.” 2000.

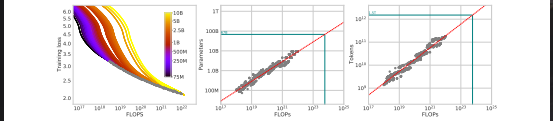


Figure 2 | **Training curve envelope.** On the **left** we show all of our different runs. We launched a range of model sizes going from 70M to 10B, each for four different cosine cycle lengths. From these curves, we extracted the envelope of minimal loss per FLOP, and we used these points to estimate the optimal model size (**center**) for a given compute budget and the optimal number of training tokens (**right**). In green, we show projections of optimal model size and training token count based on the number of FLOPs used to train *Gopher* (5.76×10^{23}).

Hoffmann et al. “Training Compute-Optimal Large Language Models.” 2022.

- 在大语言模型的预训练中，观察预训练效果最重要的指标是模型的训练loss，只要训练的loss在不断降低，我们就可以有信心这个模型在不断地学到新的知识
- 研究人员发现，模型的训练loss跟模型的参数规模、模型的训练数据大小、模型训练所使用的算力几乎都呈现某种线性关系（对数坐标下），因此，可以根据模型的参数规模、训练数据大学和训练所使用的算力来预测模型的loss，这就是所谓的Scaling Law
- 另有研究人员发现，在给定的有限算力（模型规模乘以训练量）下，存在一个最优的模型规模，并不是模型规模越大越好。太大的模型可能因为训练不充分反而达不到最优的效果。
- 直到现在，Scaling Law还没有失效，似乎预示着“大力出奇迹”的趋势还远没有停止，更强大的AI能力还有待涌现，令人无限期待，这也是大模型的竞争日趋激烈的原因。

大语言模型能力的划分

模型先天能力

由硬件性能、模型架构、规格参数等限定的能力

类比于人的先天能力，即经过亿万年生物进化获得的能力

- 度量指标：
  - 词表大小
  - 总参数量
    - 模型宽度（表示向量维度）
    - 模型深度（Transformer层数）
  - 算力消耗
  - 序列长度
  - 训练并发度、训练精度（loss）、训练速度
  - 推理并发度、推理速度（延迟）

模型后天能力

模型确定以后，通过数据训练、微调和应用获得的能力

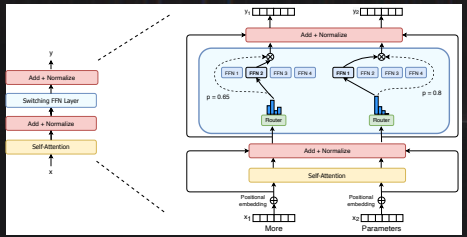
类比于人的后天能力，即通过教育和在社会中学习到的能力

- 度量指标：
  - 训练数据规模和质量
  - 语言能力、知识能力
  - 上下文学习能力、指令遵从能力
  - 数学能力、代码能力、工具使用能力
  - 推理能力、规划能力
  - 记忆能力、学习能力
  - 多模态能力
  - 行为能力

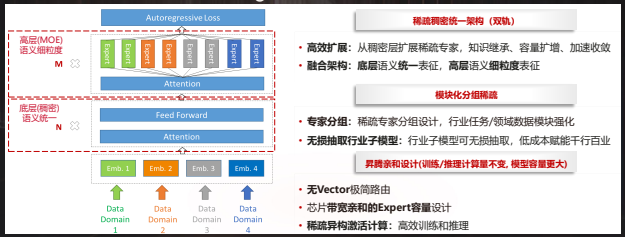
稀疏化前向网络：同等算力支持更大模型

- 挑战问题：每一步推理均需激活所有参数，不合理
- 解决方案：FFN中间层结点分组，每一个token在FFN层只激活一部分神经元

Switch Transformers. 2021.01



Pangu- Σ . 2023.03

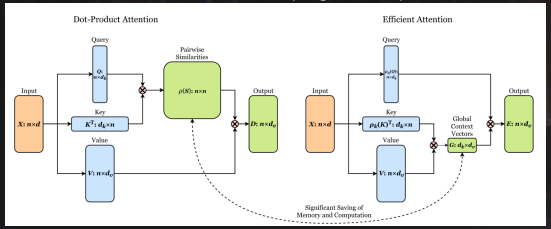


- 不久有消息披露GPT-4是采用了16xMLP111B+Attention55B结构
- 最近Mistral AI发布了开源的MoE方案以及相应的Mixtral8x7B模型，引起了更多关注
- 预计以后MoE将成为超大语言模型的标配

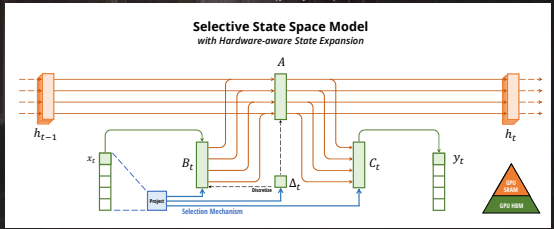
类RNN注意力机制：同等内存支持更长序列

- 挑战问题：长序列处理是复杂问题的关键。而Transformer的注意力层的推理时间和内存消耗均与序列长度的平方成正比
- 解决方案：修改注意力机制，降低计算复杂度

Linear Attention (img source)



Mamba(paper)

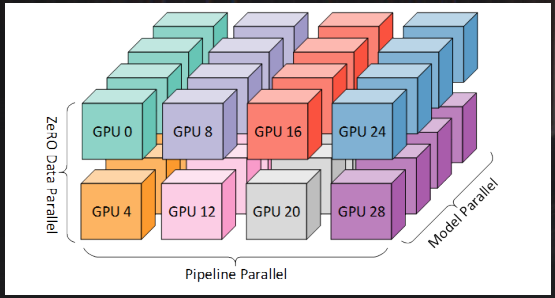


- 这个方向上研究工作非常多，近期比较有影响的工作包括RWKV、RetNet、Mamba等。
- 这些方法可以达到大大降低注意力机制计算复杂度的目的，其中一些方法在较小规模的模型上也能取得很好的性能，但都还没有在较大（100亿参数以上）的模型上得到验证。

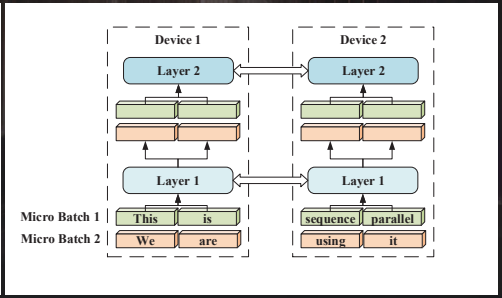
并行训练

- 挑战问题：多机并行训练的时候如果分配在各计算节点之间分配模型参数和数据，以达到最优的训练效果
- 解决方案：从数据、模型、算子、序列长度等多个维度对计算结点进行划分调度

3D并行（数据并行、模型并行、流水线并行）



序列并行

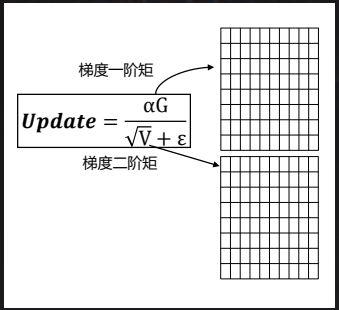


- 高效并行训练方法使得大模型训练突破的单机算力的限制，使得训练越来越大的模型成为可能
- 但计算机结点之间的通信成为新的瓶颈，芯片厂商开始推出越来越大的计算集群

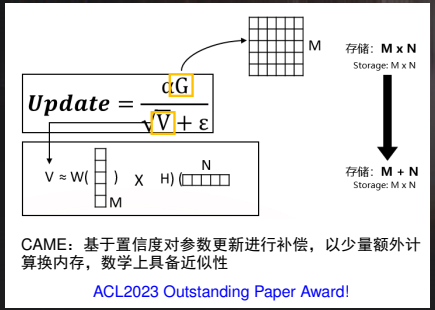
训练优化器的改进

- 挑战问题：常用的Adam优化器需存储所有参数的一阶矩和二阶矩，占用内存是参数本身的两倍。
- 解决方案：通过矩阵分解和基于置信度的参数补偿，在不影响精度的情况下，减少参数一半。

Adam优化器



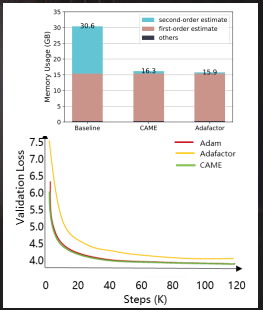
CAME：Adafactor低秩矩阵分解+参数补偿



CAME：基于置信度对参数更新进行补偿，以少量额外计算换内存，数学上具备近似性

ACL2023 Outstanding Paper Award!

实验结果

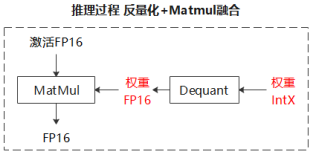


- 高效并行训练方法使得大模型训练突破的单机算力的限制，使得训练越来越大的模型成为可能
- 但计算机结点之间的通信成为新的瓶颈，芯片厂商开始推出越来越大的计算集群

推理优化

- 大模型推理挑战：生成模型参数量大，推理慢，占用内存高，端到端推理成本高。
  - 挑战一：生成模型直接使用典型的量化算法会导致严重的精度下降
  - 挑战二：推理内存占用大：1）模型参数： 1750亿模型占用350GB内存；2）KV Cache：显存占用和序列长度n成正比，1750亿模型 4k长度占用576G
- 解决方案：

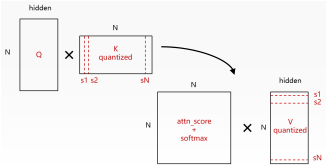
低比特量化



4/8-bit权重量化算法QuantGPT+高效反量化算子：模型内存降低2-4倍，推理加速15-30%，单卡可推380亿模型

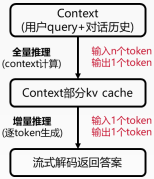
ACL2022 Outstanding Paper Award!

KV Cache量化



kv cache 8-bit量化之后内存占用降低1倍

分离部署+动态batch



动态batch：解码完成的样本提前退出，及时补进新的样本

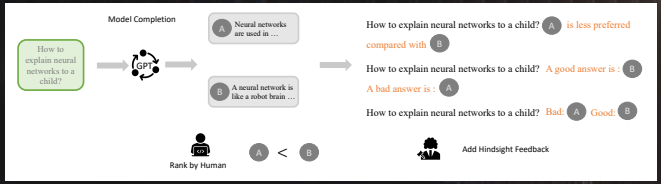
全量推理与增量推理分离部署，吞吐提升100%。

大语言模型的先天能力的提升方法小结

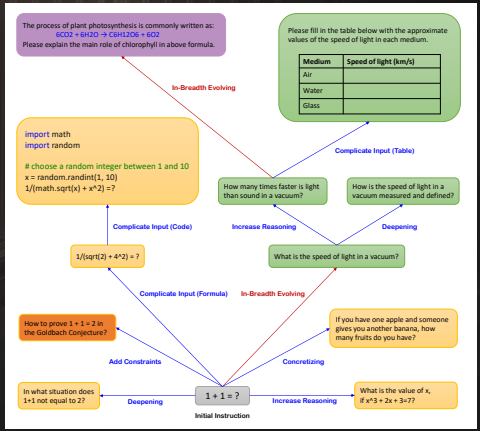
- Tokenizer
- 注意力机制的改进
  - 稀疏注意力：句首注意力、局部注意力、随机注意力、层次注意力、辅助tokens
  - 线性注意力：Linear Transformer, Performer, RWKV, RetNet
  - 状态空间模型：Mamba
  - 注意力内存优化：FlashAttention
  - 新型位置编码：RoPE
- 前向神经网络的改进
  - 混合专家系统（MoE）
  - 引入相似度查询：LookupFFN
- 非Transformer网络架构：扩散模型
- 训练优化：并行训练、新型优化器、参数量化、异构训练优化、增量训练
- 推理优化：全量增量推理分离部署、参数量化、KV Cache量化、投机推理

精心构造指令数据

Chain of Hindsight, arXiv.2302.02676.



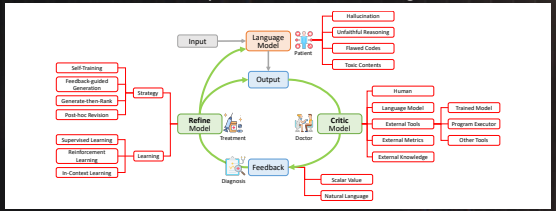
WizardLM, arXiv.2304.12244.



- 通过精心构造的指令数据，可以让模型学习到语言之间的细微区别。
- 通过系统性构造课程学习指令数据，可以让模型学到复杂的逻辑表达方式。

模型的自我评价、自我改进和自我提升

Self-critique and self-correcting



- 通过适当的微调，模型可以学习到多维度自我评价的能力。
- 通过自我评价，模型可以对自己生成的结果进行改进，生成更好的结果。
- 通过自我评价和自我改进，可以自动生成大量数据，使用这些数据进行微调可以进一步提升模型性能。
- 通过多轮自我提升迭代，可以大幅度提升模型的性能。

Self-refinement

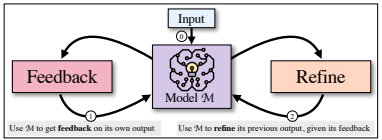


Figure 1: Given an input ((0)), SELF-REFINE starts by generating an output and passing it back to the same model $\mathcal{M}$ to get feedback ((1)). The feedback is passed back to $\mathcal{M}$, which refines the previously generated output ((2)). Steps ((1)) and ((2)) iterate until a stopping condition is met. SELF-REFINE is instantiated with a language model such as GPT-3.5 and does not involve human assistance.

SELF: iteractive self-improving

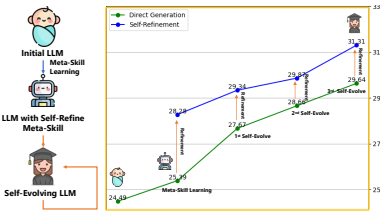
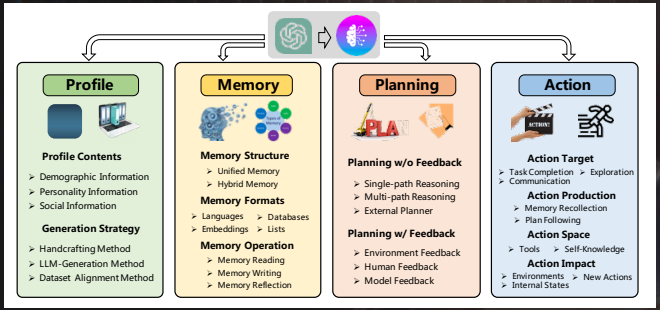


Figure 1: Evolutionary Journey of SELF: An initial LLM progressively evolve to a more advanced LLM equipped with a self-refinement meta-skill. By continual iterations (1st, 2nd, 3rd) of self-evolution, the LLM progresses in capability (24.49% to 31.31%) on GSM8K.

大语言模型驱动的智能体 LLM Agent

A Survey on Large Language Model Based Autonomous Agents. arXiv.2308.11432.



智能体区别于普通AI应用的主要特点在于：

- 智能体能够感知环境并作出决策。
- 智能体能够通过行为影响环境和改变环境。
- 智能体通过行为改变环境以后，这种改变又可以被智能体再次感知，形成闭环。
- 智能体的决策通常需要引入强化学习方法。

大语言模型驱动的智能体(LLM Agent)与传统AI智能体的区别：

- LLM Agent的状态不仅仅是向量，而且也是语言，可解释性非常好。
- LLM Agent的行为可以表示为任何复杂的函数调用等符号操作。
- LLM Agent的决策由强大的LLM提供支持。

经验的总结和积累：Voyager

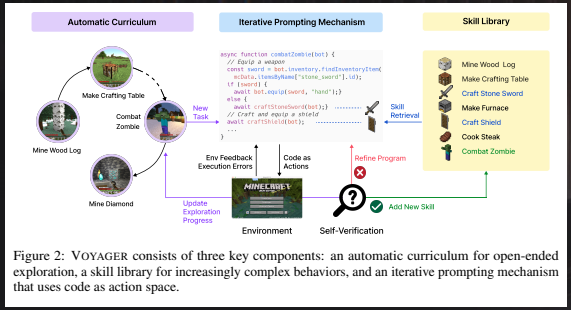
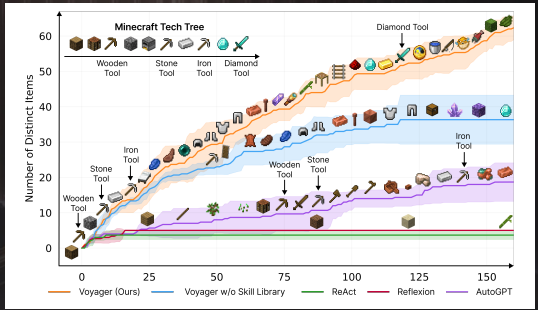
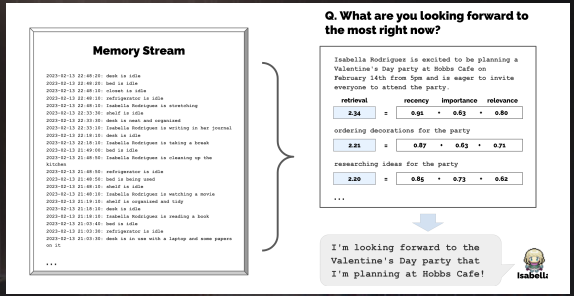


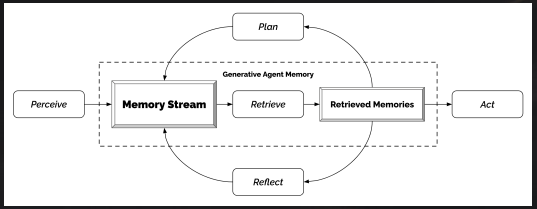
Figure 2: VOYAGER consists of three key components: an automatic curriculum for open-ended exploration, a skill library for increasingly complex behaviors, and an iterative prompting mechanism that uses code as action space.



Wang, et al. “Voyager: An Open-Ended Embodied Agent with Large Language Models.” arXiv.2305.16291.

多智能体涌现社会化行为：虚拟小镇





- 引入了基于时间的被动记忆模块。
- 决策是由大语言模型做根据记忆作出的，没有目的性。
- 多智能体之间发生了社会化行为。
- 多智能体前景展望：
  - 多智能体能否自发产生分工和合作行为，而非依靠事先指定的人设？
  - 多智能体合作可否涌现出远超单智能体的更强大的智能行为？

Park, et al. “Generative Agents: Interactive Simulacra of Human Behavior.” arXiv.2304.03442.

大语言模型的后天能力的提升小结

- 预训练：数据清洗、数据配比、数据加锁
- 指令微调：指令数据构造、课程设计、人设区隔
- 强化学习训练：RLHF（PPO/DPO）、RLAIF、自我提升、超级对齐、Q*
- 增强增强：WebGPT、RAG、向量数据库
- 工具调用：Code Interpreter、Plug-ins
- 单智能体：推理规划、思维链、路径探索、经验记忆、知识总结
- 多智能体：协作、辩论、教学、社会行为
- 多模态：音频、图像、3D、视频
- 行为交互（具身智能）

幻象问题

- 现有神经网络本身无法避免幻象问题：
  - 神经网络的知识都存储在参数之中，参数本身是无法区分事实和幻象的。
  - 神经网络生成的文本或图像也无法区分事实和幻象。
- 更大的模型（如GPT-4）可以更好地对数据建模，有助于减少幻象。
- 引入外部知识（如RAG），可以很好地减少幻象。
- 人类也有幻象（儿童、病人、梦境、文学创作），幻象不一定都是坏事。
- 一种可能的消除幻象的方案，是在模型内部引入事实性判断模块。
- 在特定应用领域，只要能把幻象减少到足够低，就可以满足需求，并不一定需要彻底消除幻象。

超人智能和毁灭人类的风险

- AI能力将在越来越多的专业领域内超过大部分普通人、甚至超过一般的专家，成为超人智能（Superhuman Intelligence）。
- AI在日常生活中超过人类，还有很长的路要走。
- AI毁灭人类的可能性：
  - AI没有毁灭人类的意图（有能力并不等于有意图）。
  - AI可能无意中毁灭人类：回形针思想实验。
  - 回形针思想实验的问题（个人意见）：
    - 资源的授权；
    - 从失败中恢复的意志。

对未来社会的影响

- 智能和能源的成本将趋近于零。
- AI将像水、电、无线通信、互联网一样，成为每个人日常生活不可或缺的基础设施。
- AI驱动的科学研究（AI4Science）将带来科学的革命，大大加速科学进步的速度。
- AI将对人类社会的组织形态造成巨大冲击：
  - 高智力岗位将继续存在，但门槛将大大提高。
  - 一些低智力重复劳动的工作将消失。
  - 一些机器无法取代的体力工作和具有较高情绪价值的服务型工作将长期存在并可能成为大部分人的工作。
  - AI将使人类更少工作，而有更多投入时间进行学习和娱乐。
  - 普遍基本收入（Universal Basic Income，UBI）将无可避免。

Summary

大模型技术概览及总体趋势

大语言模型先天能力发展趋势

大语言模型后天能力发展趋势

大语言模型问题、风险和对社会的影响

总结

2023 JUEJIN 2023 TECH TALK
掘金年度技术演讲

Thank you!

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

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

HUAWEI