Large Language Models - Technology Status, Trends and Impacts

刘群 LIU Qun

华为诺亚方舟实验室 Huawei Noah's Ark Lab

A Guest Talk to Hong Kong Metropolitan University

2024-03-07, Online

NOAH'S ARK LAB

HUAWEI

Content

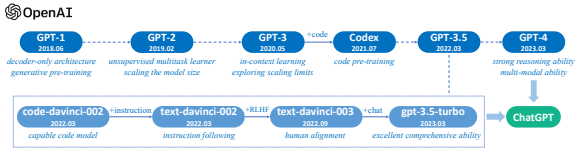
Content

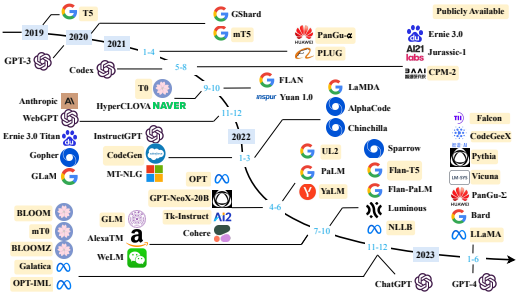
What are Large Language Models (LLMs)?

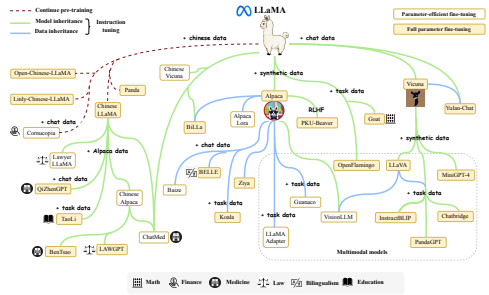
- Large Language Models (LLMs) are statistical language models with huge number (normally more than 1 billion) of parameters.
- LLMs are originally language models but can be extended to multimodal models which can process audio, image and video data.
- LLMs are also known as foundation models, although there are subtle differences between them.
- Typical LLMs include GPT-3/3.5/4, ChatGPT, Claude, LLaMA, Gemini etc.

Large Languege Models: an overview

- 2023 is the year of explosion of LLMs
- 2023 is also called the meta-year of AGI
- LLMs have had profoundly impacted AI
- LLMs will further profoundly impact our society







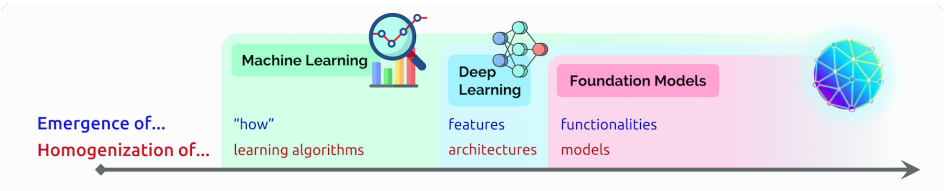
Zhao, et al. “A Survey of Large Language Models.” arXiv2303.18223

HUAWEI NOAH'S ARK LAB

Emergence and homogenization of foundation models

The significance of foundation models can be summarized with two words: emergence and homogenization.

- Emergence means that the behavior of a system is implicitly induced rather than explicitly constructed; it is both the source of scientific excitement and anxiety about unanticipated consequences.
- Homogenization indicates the consolidation of methodologies for building machine learning systems across a wide range of applications; it provides strong leverage towards many tasks but also creates single points of failure.



	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)

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

Homogenization: Pre-trained LMs vs. LLMs

	Pre-trained Language Models (PLMs)	Large Language Models (LLMs)
Typical Models	ELMo, BERT, GPT	GPT-2, GPT-3
Model Architectures	BiLSTM, Transformer	Transformer
	Encoder, Encoder-decoder, Decoder	Decoder
Attention Directions	Bidirectional、Unidirectional	Unidirectional
Training Methods	Mask & Predict Autoregressive Generation	Autoregressive Generation
Task Types	NLU	NLU & NLG
Model Sizes	0.1-1B parameters	1Billion-xTrillion parameters
Applying Methods	Fine-tuning	Prompting & Fine-tuning & RLHF
Emergent Abilities	Inductive Transfer Learning	Zero-shot Learning Few-shot/In-context Learning Chain-of-Thought

Ability emergence in LLMs

In-context Learning (zero/few shot learning)

Zero-shot

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

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

One-shot

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

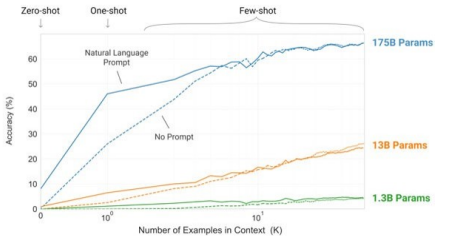
```
1 Translate English to French:        ← task description
2 sea otter => loutre de mer          ← example
3 cheese =>                           ← prompt
```

Few-shot

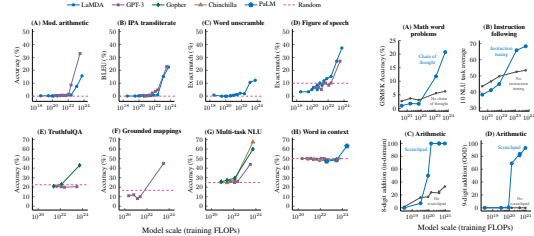
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

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

Emergence of in-context learning



Emergence of other abilities



HUAWEI NOAH'S ARK LAB

LLM training: the Scaling Law

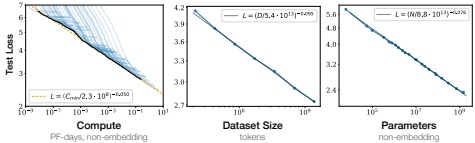


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when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

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

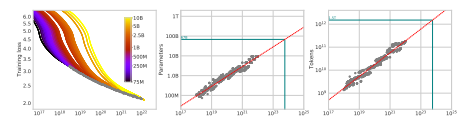


Figure 2 | **Training curve envelope.** On the **left** we show all of our different runs. We launched a range of model sizes going from 70M to 10B, each for four different cosine cycle lengths. From these curves, we extracted the envelope of minimal loss per FLOP, and we used these points to estimate 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.

- The Scaling Law predicts that the performance of LLMs will increase along with the increasing of the model size, the training data amount, as well as the consumption of computing power in training.
- This encourage the industries to consistently persue larger models and more training data, to obtain a more powerful AI.
- It is estimated that when the number of parameters of LLMs reaches 100 trillion, which is comparible with the number of synapses of human brains, the AGI will achieved.

HUAWEI NOAH'S ARK LAB

LLM abilities: a classification

Bulit-in Abilities

Abilities built-in and limited by hardware specifications, hardware performance, and the model architecture

Analogy to the innate abilities of human beings, obtained through hundreds of millions of years of biological evolution

- Sparse(MoE) or dense, number of experts
- Total Parameter Number, number of attention heads
- Model Width (Dimension of the representative vectors)
- Model Depth (Number of layers of the Transformer model)
- Computing power consumption in training (FLOPS)
- Sequence length, vocabulary size
- Training parallism, training loss, training speed
- Inference parallism, inference speed (delay)

Learned Abilities

Abilities obtained through data training, fine-tuning, and application given a specific model

Analogy to acquired human abilities, i.e., abilities learned through education and from the society

- Quality and quantity of the tranining data
- Language abilities, knowledge abilities
- In-context learning abilities, instruction-following abilities
- Math abilities, coding abilities, tools-using abilities
- Reasonging abilities, planning abilities
- Memorization abilities, learning abilities
- Multi-modal abilities
- Action abilities

Content

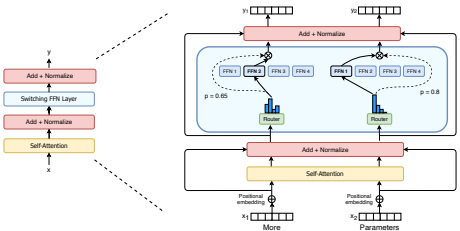
Overview of technologies for improving LLM built-in abilities

- Improvement of tokenizer
- Improvement of attention mechanism
  - Sparse attention: SparseTransformer, BigBird, LocalAttention, ...
  - Linear attention: Performer, RWKV, RetNet
  - Space state models: Mamba
  - Memory usage optimization: FlashAttention
  - Novel positional encoding: RoPE
- Improvement of FFNs
  - Sparse FFN: MoE
  - Replacing calculation with retrieval: LookupFFN
- Non-transformer Models: Diffusison models
- Training improvement: parallism, efficient optimizer, quantization, heterogenerious training, incremental training
- Inference improvement: seperated deployment, parameter quantization, KV cache quantization, speculative decoding

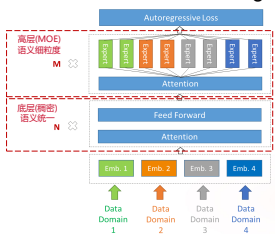
Sparse FFNs: support larger models with the same computing power

- Challenge: all the parameters are activated at each step of inference, which is too expensive.
- Solution: Mixture-of-Experts (MoE): the internal nodes of FFN are grouped into experts, while only part of groups are activated at each time.

Switch Transformers. 2021.01



Pangu- Σ . 2023.03



稀疏稠密统一架构（双轨）

- **高效扩展**：从稠密层扩展稀疏专家，知识继承、容量扩增、加速收敛
- **融合架构**：**底层**语义**统一**表征，**高层**语义**细粒度**表征

模块化分组稀疏

- **专家分组**：稀疏专家分组设计，行业任务/领域数据模块强化
- **无损抽取行业子模型**：行业子模型可无损抽取，低成本赋能千行百业

昇腾亲和设计(训练/推理计算量不变，模型容量更大)

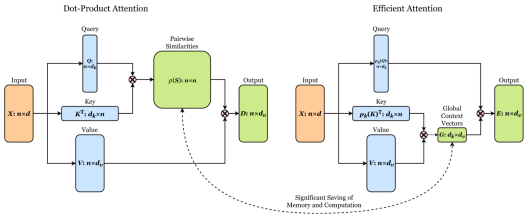
- **无Vector**极简路由
- 芯片**带宽亲和的Expert容量**设计
- **稀疏异构激活计算**：高效训练和推理

- According unofficially disclosed information, GPT-4 adopts an MoE architecture.
- Mistral AI released the source codes of its MoE model Mixtral 8x7B.
- It is expected that the MoE architecture will be popular in the futhure LLMs.

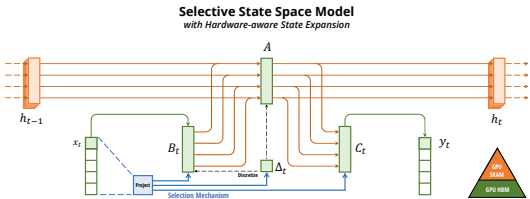
RNN-like attentions: support longer context with the same memory size

- Challenge: Long sequence processing is crucial to complex problems. The inference time and memory consumption of the attention layer of Transformer are proportional to the square of the sequence length.
- Solution: Modify the attention mechanism to reduce the computing complexity.

Linear Attention (img source)



Mamba(paper)

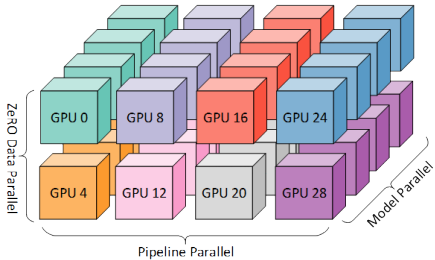


- There are a lot of work in this direction, including recently released work like RWKV, RetNet, Mamba, etc.
- These methods can reduce the computational complexity of the attention mechanism, and some of them can also achieve good performance on smaller-scale models, but none of them has been verified on larger (more than 10 billion parameters) models.

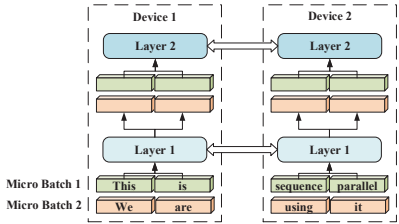
Training parallism

- Challenge: How to allocate model parameters and data among computing units during parallel training to achieve the optimal training effect?
- Solution: Partition the computing units from multiple dimensions, such as data, model layers, operators, and sequence length.

3D parallism (data, model, pipeline)



sequence parallism



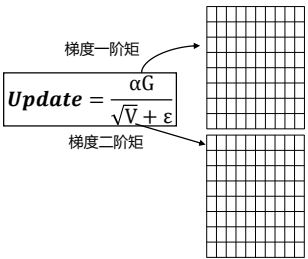
- The development of parallel training methods makes it possible to train larger and larger models.
- However the communication between computer units became the new bottleneck, and manufacturers are producing large computing clusters, which support high-speed communication between computer units inside a cluster.

HUAWEI NOAH'S ARK LAB

Improvement of training optimizers

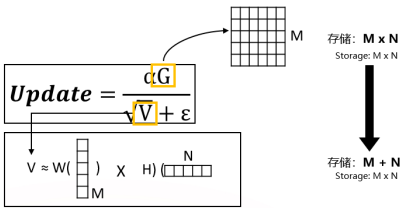
- Challenge: The commonly used Adam optimizer stores the first-order and second-order moments of all parameters, which occupies twice the memory of the parameters.
- Solution: By matrix decomposition and parameter compensation based on confidence, the memory usage are reduced by half without affecting the accuracy.

Adam & Adafactor Optimizer



Adafactor: Low-rank matrix factorization
Precision decreased.

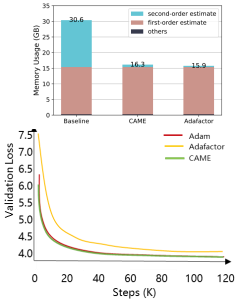
CAME Optimizer



CAME: Adafactor + Parameter Compensation
Parameter Compensation based on confidience level.
Exchange memory with a little calculation.
Mathematical approximation.

ACL2023 Outstanding Paper Award!

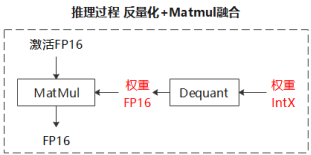
Results



Improvement of inference

- Challenges of inference: large number of parameters, slow inference, high memory usage, and high cost of end-to-end inference.
  - Challenge 1: Existing quantization will cause severe accuracy deterioration.
  - Challenge 2: High memory usage: 1) Parameters: 175B model -> 350GB memory
    2) KV Cache: linear to seq.len. 175B model + 4K context -> 576GB memory
- Solution:

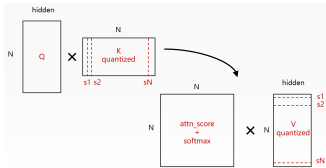
Low-bit quantization



- 4/8-bit QuantGPT
- Efficient dequantization operator
- 2-4x memory reduction
- Inference acceleration: 15-30%
- 38B model inference in a single card

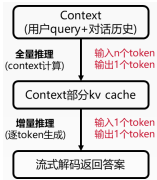
ACL2022 Outstanding Paper Award!

KV Cache Quantization



- kv cache 8-bit quantization
- 1x memory reduction

Separate deployment+Dynamic batch



- Dynamic batch: replacing completed samples with new samples instantly
- Separate deployment of full inference and incremental inference, improving throughput by 100%

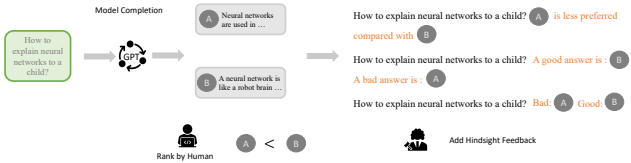
Content

Overview of technologies for improving LLM learned abilities

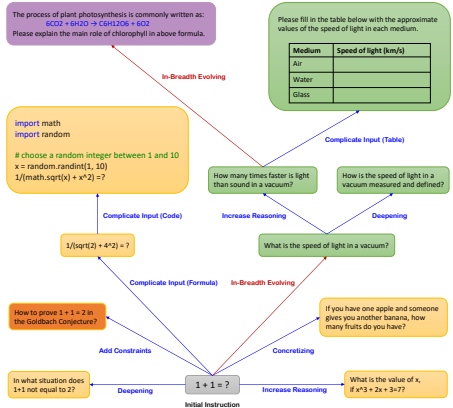
- Pre-training: data cleaning, data proportioning, data safe-guarding
- Instruction tuning: Instructional data construction, curriculum design, and persona segregation
- RL training: RLHF (PPO/DPO), RLAIF, self-improvement, super-alignment, Q*
- Retrieval-augmentation: WebGPT, RAG, vector database
- Tool using: Code Interpreter, plug-ins
- AI agent: Reasoning, planning, chain-of-thought, path exploration, experience memorization, knowledge summarization
- Multi-agent: Collaboration, debate, teaching, social behavior
- Multi-modal: audio, image, 3D, video
- Behavior and interaction: embodied AI

Elaborated Instruction Data

Chain of Hindsight, arXiv.2302.02676.



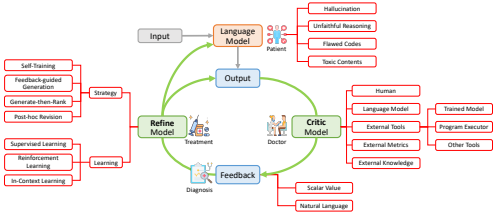
WizardLM, arXiv.2304.12244.



- With carefully curated instruction data, the model can learn the subtle differences between languages.
- By systematically constructing course learning instruction data, the model can learn complex logical expressions.

Self-critique, self-correcting and self-improving of LLMs

Self-critique and self-correcting



- With appropriate fine-tuning, the model can learn multi-dimensional self-evaluation.
- By using self-evaluation, the model can improve the results it generates and generate better results.
- By using self-evaluation and self-correction, a large amount of data can be automatically generated, and the model performance can be further improved by using this data for fine tuning.
- Through multiple self-improvement iteration, the performance of the model can be greatly improved.

Self-refinement

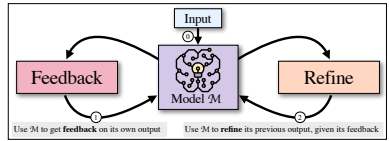


Figure 1: Given an input (⓪), SELF-REFINE starts by generating an output and passing it back to the same model $\mathcal{M}$ to get feedback (①). The feedback is passed back to $\mathcal{M}$, which refines the previously generated output (②). Steps (①) and (②) iterate until a stopping condition is met.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

SELF: iterative self-improving

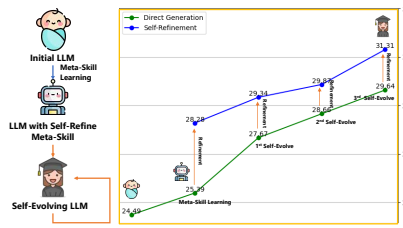
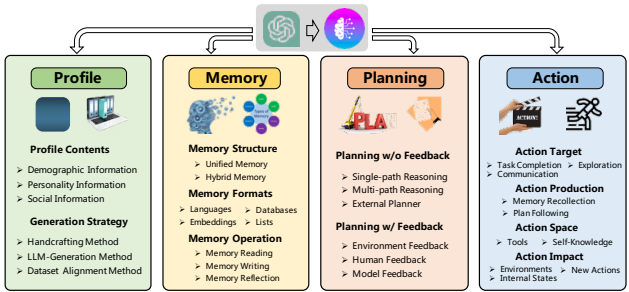


Figure 1: Evolutionary Journey of SELF: An initial LLM progressively evolve to a more advanced LLM equipped with a self-refinement meta-skill. By continual iterations (1st, 2nd, 3rd) of self-evolution, the LLM progresses in capability (24.49% to 31.31%) on GSM8K.

HUAWEI NOAH'S ARK LAB

LLM-driven AI agents

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



Difference between AI agents and common AI applications:

- Agents are able to perceive the environment and make decisions.
- Agents can influence and change the environment through their behavior.
- Agents can perceive the changes of the environments caused by their own behaviour, which form a close loop.
- The learning of the decision-making mechanism of agents usually involve reinforcement learning.

Differences between LLM-driven agents and traditional AI agents:

- The states of LLM Agent are represented not only with vectors, but also in languages, which is interpretable.
- The behavior of LLM agents can be represented as any complex symbolic operation such as function calls.
- The LLM Agent's decision is supported by a strong LLM.

Summarization and accumulation of experience: Voyager

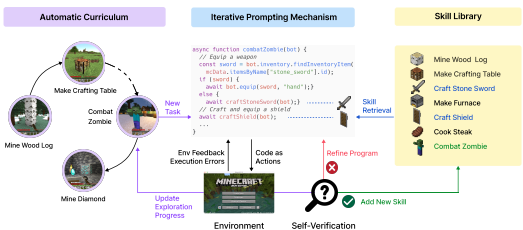
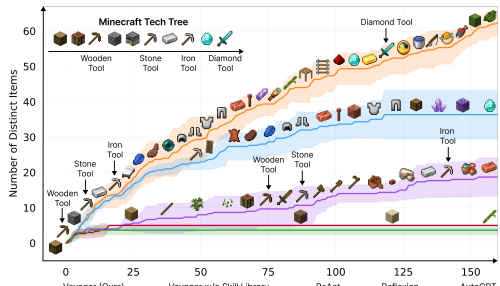
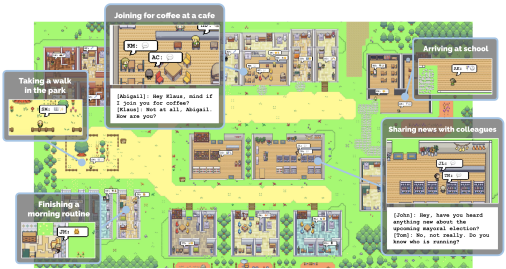


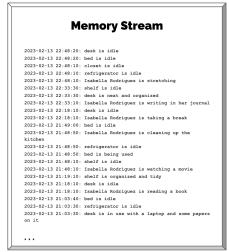
Figure 2: VOYAGER consists of three key components: an automatic curriculum for open-ended exploration, a skill library for increasingly complex behaviors, and an iterative prompting mechanism that uses code as action space.



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

Emergent social behavior from multi-agent interaction: Smallville





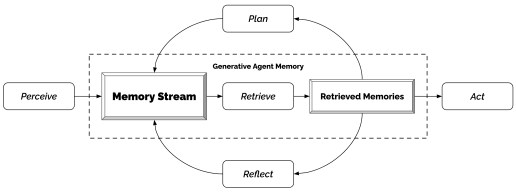
Q. What are you looking forward to the most right now?

Isabella Rodriguez is excited to be planning a Valentine's Day party at Hobbs Cafe on February 14th from 5pm and is eager to invite everyone to attend the party.

	retrieval		recency		importance		relevance
ordering decorations for the party	2.34	=	0.91	•	0.63	•	0.80
researching ideas for the party	2.21	=	0.87	•	0.63	•	0.71
...	2.20	=	0.85	•	0.73	•	0.62

I'm looking forward to the Valentine's Day party that I'm planning at Hobbs Cafe!

Isabella



- Introduce time-based passive memory.
- Decisions are made by LLMs according to memory, without purposes.
- Social behavior emergents among multi-agents.
- Potential future development of multi-agent:
  - Can division of labor and cooperative behavior emergent among multple agents, rather than relying on pre-specified human design?
  - Can ever more powerful intelligent behaviour emergents through collaborations between multi-agents?

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

HUAWEI NOAH'S ARK LAB

Content

Hallucination

- The current neural networks are unable to avoid hallucination inherently:
  - The knowledge of a neural network is stored in the parameters, and the parameters themselves cannot distinguish between fact and hallucination.
  - It is also not possible to distinguish between fact and hallucination in texts or images generated by neural networks.
- Larger LLMs, such as GPT-4, can better model the data and help to reduce hallucination.
- Introducing external knowledge, such as RAG, can reduce hallucination.
- Humans also have halluciations (children, patients, dreams, literary creations), and hallucinations are not always bad things.
- A possible solution to eliminate halluciation is to introduce a factual examination module inside the model.
- In specific applications, it is acceptible to reduce the halluciation to a low enough level, rather than completely eliminate the hallucination.

Superhuman intelligence: the risk of dominating/destroying humanity

- AI abilities will exceed most average person and even experts in more and more professional fields.
- AI has a long way to go before it surpasses humans in daily life.
- The possibility of AI to dominate or destroy humanity:
  - AI has no intention of dominating or destroying humanity (having the ability is not the same as having the intention).
  - AI could unintentionally destroy humanity: the paperclip thought experiment.
  - The problem with the paperclip thought experiment (my personal opinion):
    - Humans will not give up the authorization of resources;
    - AI does not have a strong will to recover from failure.

The impact of LLMs to the future of human society

- The cost of intelligence and energy will be close to zero.
- AI will become an indispensable infrastructure for everyone's daily life, just like water, electricity, wireless communication, and the Internet.
- AI-driven scientific research (AI4Science) will bring about a scientific revolution and greatly accelerate the speed of scientific progress.
- AI will have a huge impact on the organization of human society:
  - High-intelligence jobs will continue to exist, but the bar will be raised significantly.
  - Some low-intelligence and repetitive work jobs will disappear.
  - Most people will take manual jobs that machines can't replace or service jobs with high emotional value.
  - AI will allow humans to work less and spend more time learning and having fun.
  - Universal basic income (UBI) is inevitable.

Content

Summary

Trends of Large Languge Models: an high-level overview

Trends of technologies for improving LLM built-in abilities

Trends of technologies for improving LLM learned abilities

Challenges, risks and social impacts of LLMs

Summary

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