GOSIM

GOSIM AI Paris 2025

Decode DeepSeek:

the Technological Innovation and Its Influence on AI Ecosystem

Jason Li, SVP@CSDN

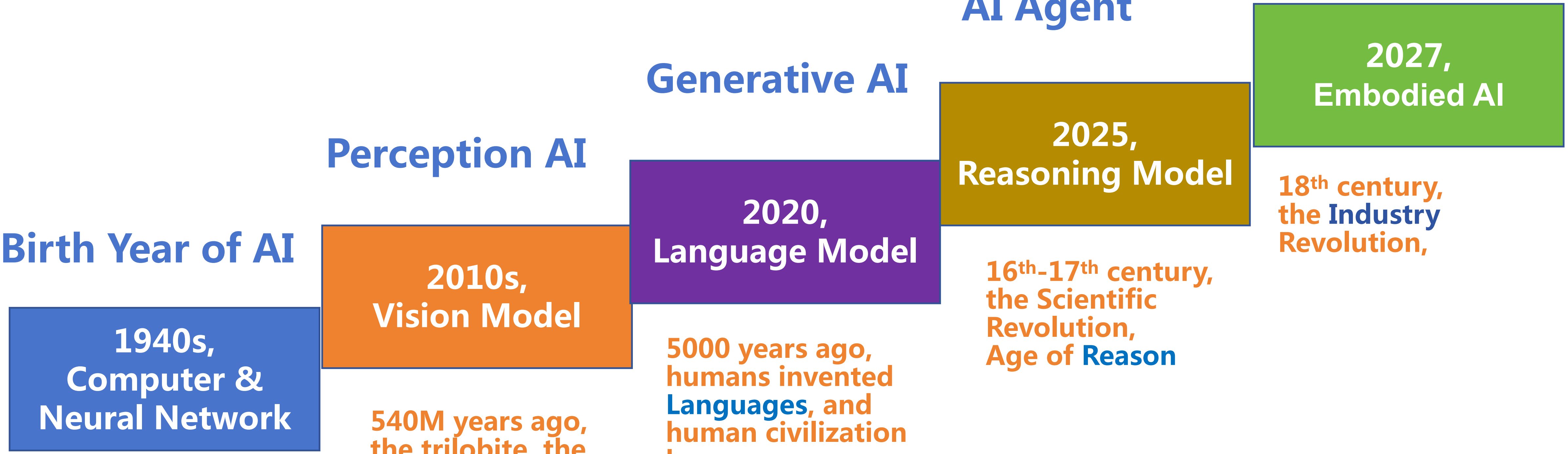
Email: lijz@csdn.net
Linkedin: https://www.linkedin.com/in/jianzhongli

GOSIM

Jason Li, Senior Vice President of CSDN, Chief Technical Expert of Boolan

Chairman of Machine Learning Summit (ML-Summit) , and member of the ISO-C++ Standard Committee. He has rich experience and in-depth research in artificial intelligence, software architecture and product innovation. In recent years, he has mainly focused on the application of artificial intelligence methods centered around Large Language Models. He proposed the "ParaShift Cube" for technological product innovation. He is also a serial entrepreneur who has founded well-known technology companies including SoftCompass, SlideIdea, and Boolan.

The development of human intelligence and AI echoes each other.



Birth Year of AI

1940s, Computer & Neural Network

550M years ago,

Caenorhabditis elegans, the first organism with a Brain & Neural Network

Perception AI

2010s, Vision Model

540M years ago, the trilobite, the earliest organism with Eyes

Generative AI

2020, Language Model

5000 years ago, humans invented Languages, and human civilization began.

AI Agent

2025, Reasoning Model

16th-17th century, the Scientific Revolution, Age of Reason

Physical AI

2027, Embodied AI

18th century, the Industry Revolution,

“I think, therefore I am”

René Descartes (1596 —1650)

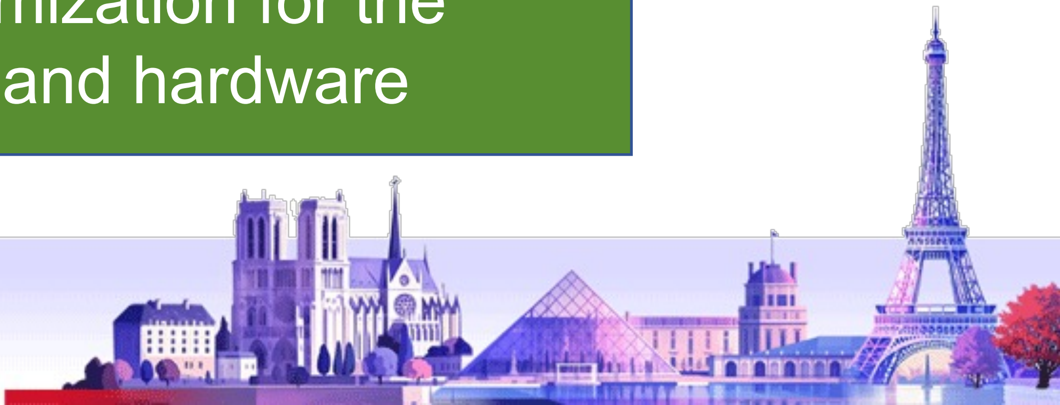
DeepSeek’s Key Technological Innovations

The three key innovations of DeepSeek

1. Open source reinforcement learning leads the paradigm shift of the inference computing.

2. The innovation of model architecture achieved through MLA and MoE

3. "Tailor-made" engineering optimization for the collaboration between software and hardware



Thinking, Fast and Slow by Daniel Kahneman

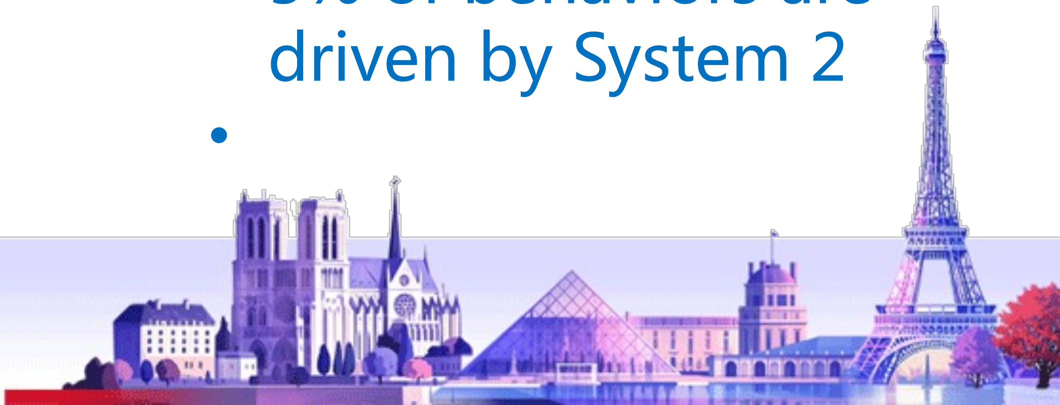
GOSIM

System 1 : Fast Thinking

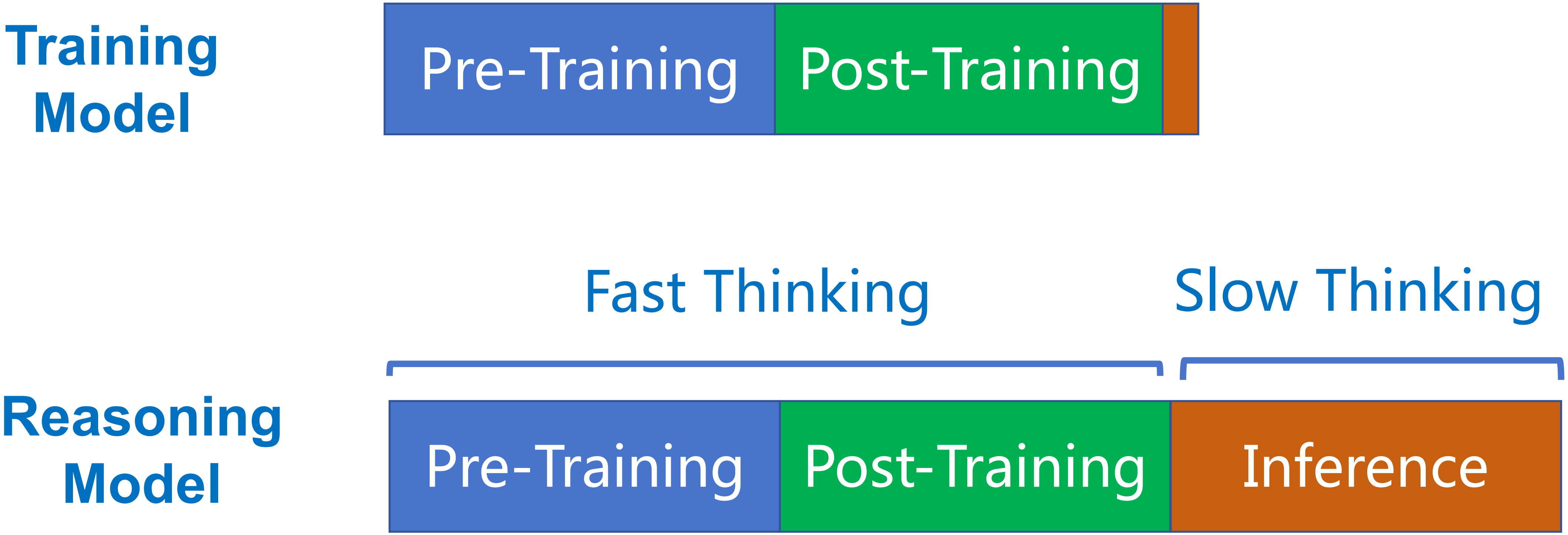
- intuitive
- unconscious
- stereotypic
- automatic
- frequent
- 95% of behaviors are driven by System 1

System 2: Slow Thinking

- logical
- conscious
- calculating
- effortful
- infrequent
- 5% of behaviors are driven by System 2



Training model VS. reasoning model



Training Model

Pre-Training

Post-Training

Fast Thinking

Slow Thinking

Reasoning Model

Pre-Training

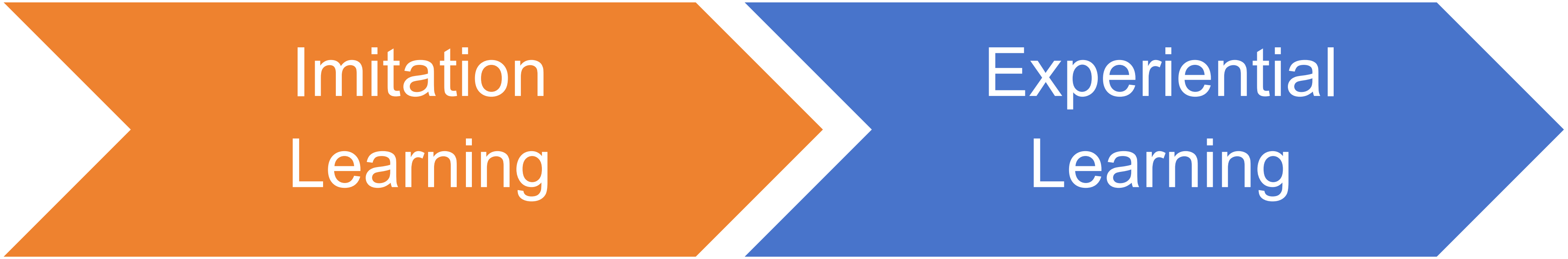
Post-Training

Inference

Reinforcement Learning drives LLM’s paradigm shift of the inference computing

- OpenAI O1 (Step, 2024) and DeepSeek R1 (Jan, 2025, **Open Source**) led LLM’s paradigm shift driven by Reinforcement Learning together.
- DeepSeek successfully achieved pure Reinforcement Learning without SFT (Supervised FineTuning) and did not rely on cold-start data.
- Reinforcement Learning generates data through a "**reward-punishment**" model and has a **scaling** effect. Chain of Thought (COT) in Pre-training does not have a scaling effect.

From imitation to experiential learning



Imitation Learning

Experiential Learning

- System 1 Thinking
- Deep Learning
- Human-Generated Data
- Echo Chamber of Human Knowledge

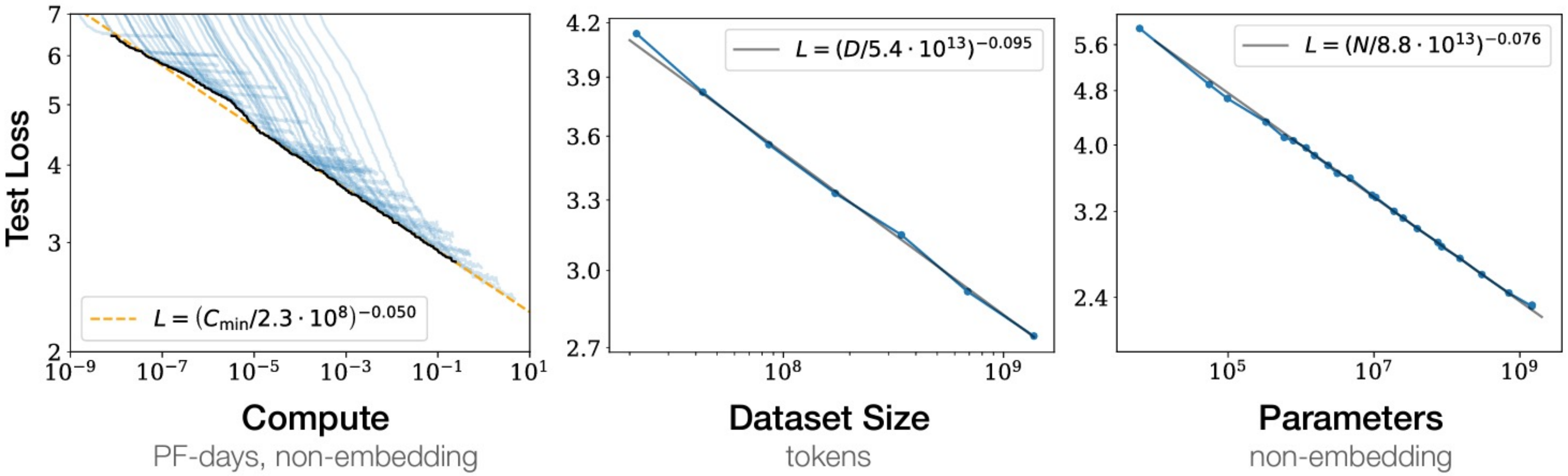
- System 2 Thinking
- Reinforcement Learning
- Machine-Generated Data
- Exploratory Innovation Beyond Human Knowledge

Scaling Law

- 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.

The Bitter Lesson by Richard Sutton

- The success of the Transformer architecture lies in its adaptation to the **Scaling**.
- Scaling Law describes how the performance of AI systems improves as the size of the training data, model parameters or computational resources increase.



Does Scaling Law End ?

No, the Scaling Law does not end.

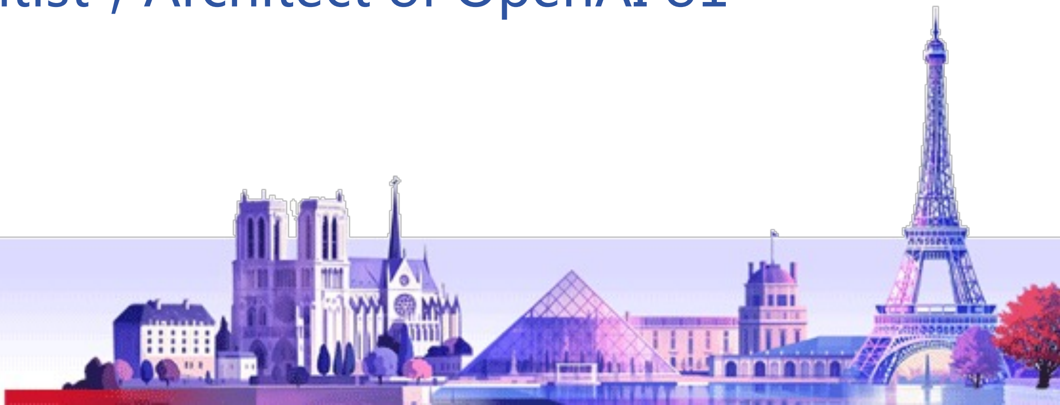
It is because the data for pre-training is nearly exhausted, so that the scaling speed in the pre-training stage slows down.

But ……

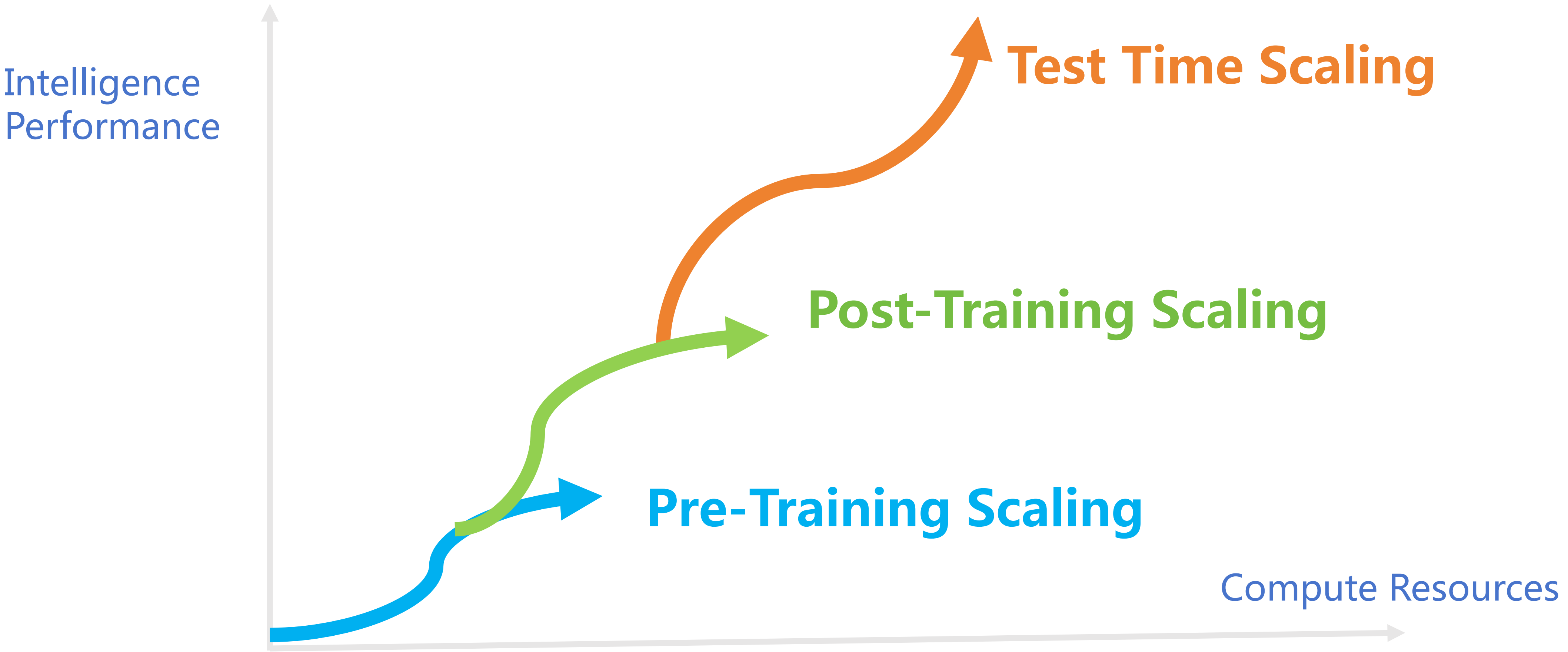
It turned out that having a bot think for just 20 seconds in a hand of poker got the same boosting performance as scaling up the model by 100,000x and training it for 100,000 times longer.”

Noam Brown

Research Scientist，Architect of OpenAI o1

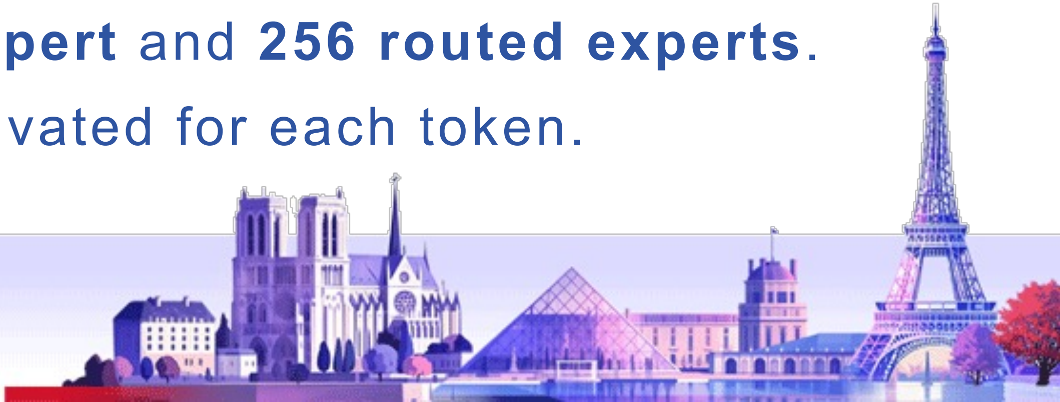


Three Phases of Scaling Law in LLM



The innovations of model architecture

- MLA (Multi-head Latent Attention) achieves **efficient inference** through significantly compressing the Key-Value (KV) cache into a latent vector.
- The MLA mechanism reduces the KV cache by 93.3%, leading to a smaller memory footprint during inference. and significantly enhances inference efficiency.
- MoE (Mixture-of-Experts), achieves **economical training** through sparse computation, enabling model to activate only a subset of its total parameters for each token. Scales to **671 billion** parameters, with **37 billion activated** per token.
- DeepSeekMoE uses finer-grained experts and isolates some experts as shared ones. Each MoE layer consists of **1 shared expert** and **256 routed experts**. Among the routed experts, 8 experts will be activated for each token.



Engineering optimization for the integration between software & hardware

- Engineering optimization strategies of software-hardware collaboration have been implemented at multiple aspects, including **computing**, **communication**, and **storage**.
- Multi-Token Prediction, DualPipe, Cross-Node Communication, FP8 quantization & hybrid parallelism, Auxiliary-loss-free load balancing, etc.
- If DeepSeek becomes the standard in the field of open-source LLM , there will be an opportunity to implement **"model-defined hardware”,** which could subvert NVIDIA's CUDA ecosystem.

DeepSeek’s Key Influences on AI Eco-System

The three key influences by DeepSeek

1. The cost of large languange models is reduced significantly，AI applications (agents) are booming.

2. DeepSeek is refactoring the AI computing eco-system in both of training and Inference stage.

3. The open source with longtermism is the cradle of innovation

Cost Matters !

Training Cost

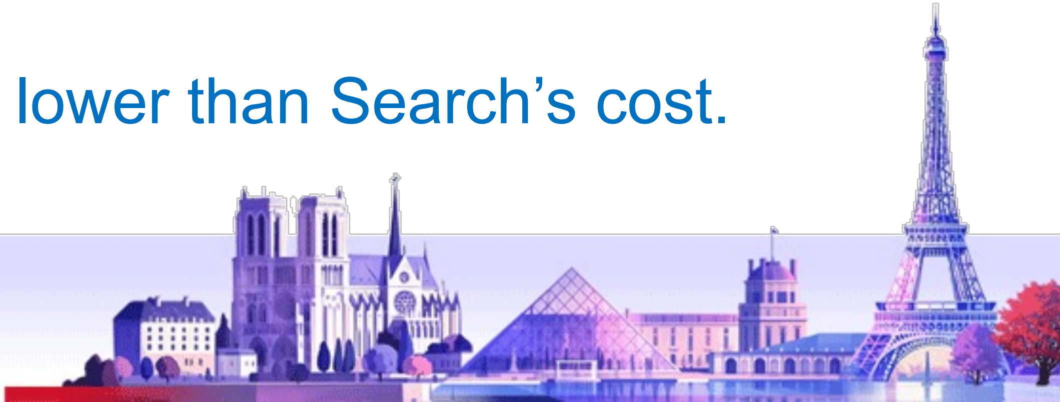
DeepSeek V3	OpenAI GPT-4	Llama 3
~5.6 M US$	78M ~ 100M US$	92M~123M US$

Inference Cost

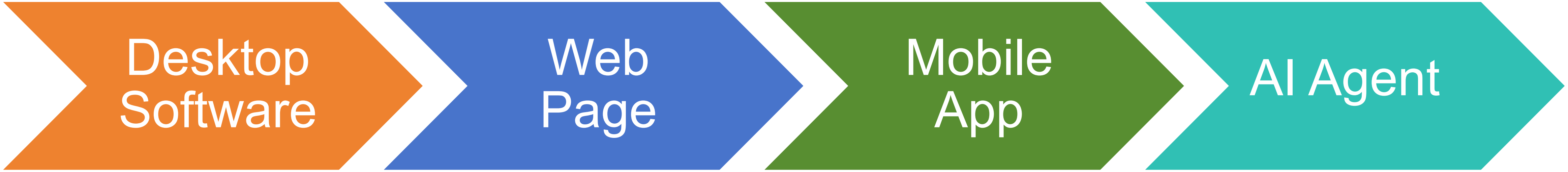
Google Search	DeepSeek R1 Reasoning	DeepSeek V3 Chat	OpenAI GPT 4.1	OpenAI O3
0.2 Cents / one time	0.2 Cents / 1000 tokens	0.025 Cents / 1000 tokens	0.8 Cents / 1000 tokens	4 Cents/ 1000 tokens

Chinese Experience: Low Cost is important to scale. Free Mode is disruptive.

AI Applications are booming when LLM’s cost is lower than Search’s cost.



What is the application form in AI Era?



Desktop Software

Web Page

Mobile App

AI Agent

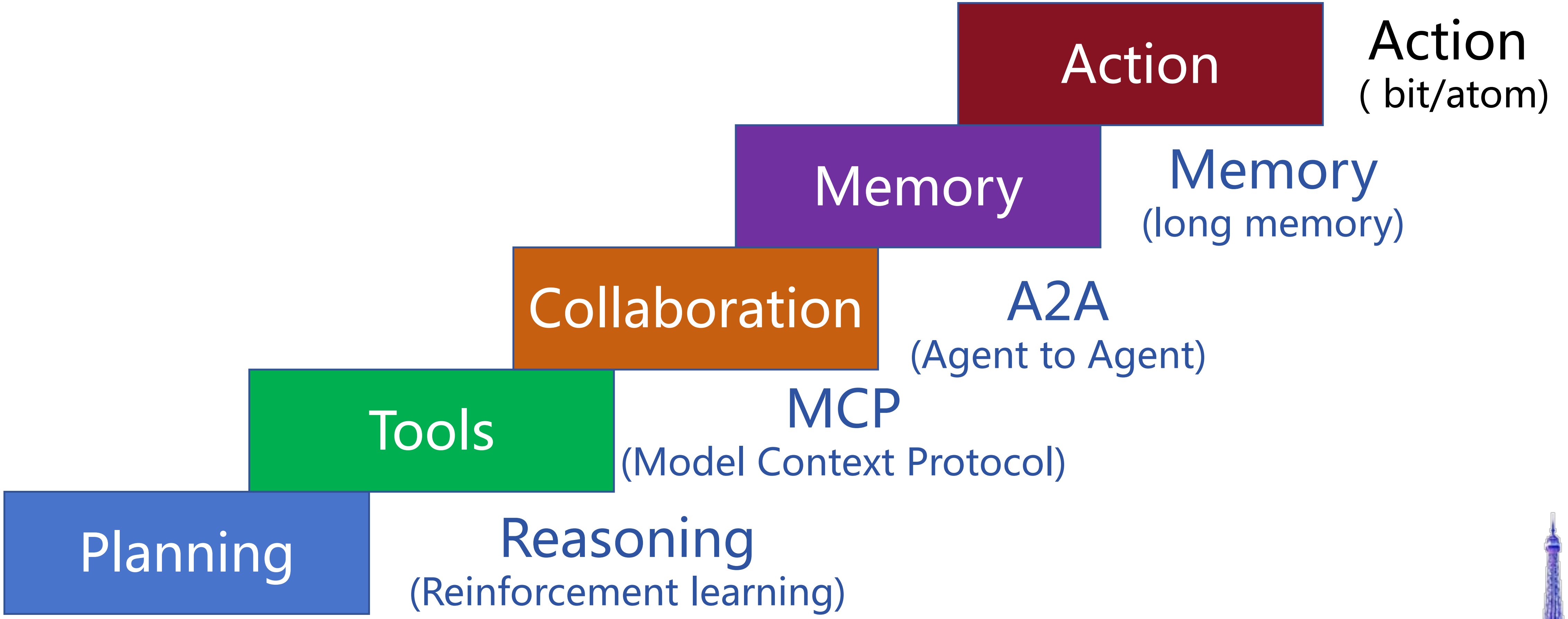
PC Era

Internet Era

Mobile Era

AI Era

The Core Capability Model of Agent



Action

(bit/atom)

Memory

(long memory)

A2A

(Agent to Agent)

MCP

(Model Context Protocol)

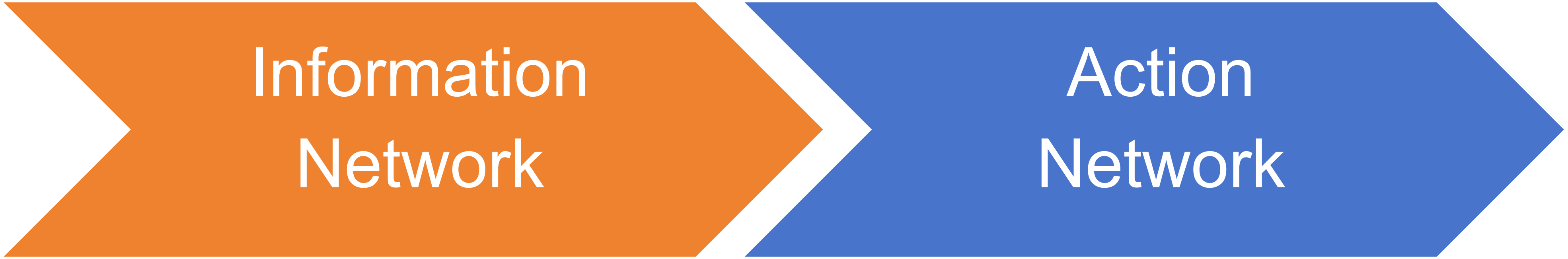
Reasoning

(Reinforcement learning)

The transformation of computing tasks

Retrieval Mode	Generative Mode	Action Mode
Access/Storage	Learn	Plan
Search	Predict	Collaborate
Compute	Create	Execute
Digitalization	**GenAI/AIGC**	**AI Agent**

Agent is transforming Internet from "Information Network" into "Action Network"



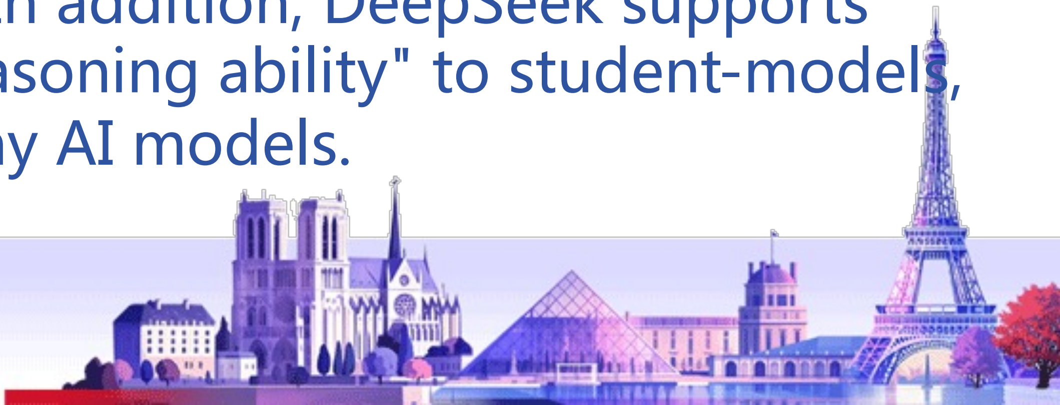
- Humans get information
- Humans make plans
- Humans take actions
- Humans get results

- Humans prompt needs
- Agents make plans
- Agents take actions
- Humans get results

DeepSeek is Refactoring the Eco-system of AI Computing

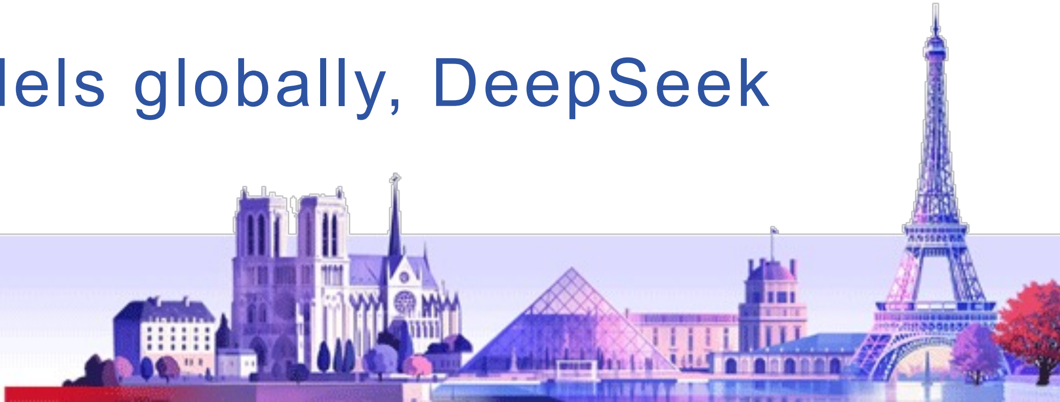
- **1、Inference stage**: As Reinforcement Learning is driving the reasoning model, the proportion of inference computing will increase rapidly. Pre-training computing (dominant advantage of NVIDIA) will no longer be the mainstream of AI computing requirements. In inference computing, there are many choices, such as Huawei's Ascend ,Google's TPU, and AWS's Inferentia. Additionally, there are also diverse inference scenarios, such as distributed computing, and edge computing.

- **2、Training stage:** Through the improvements of Transformer architecture, such as MoE and MLA, DeepSeek achieves pre-training of the same scale with 10 to 20 times less computing resources than its peers. In addition, DeepSeek supports using R1 as the teacher-model to distill the "reasoning ability" to student-models, which will also reduce the training costs of many AI models.



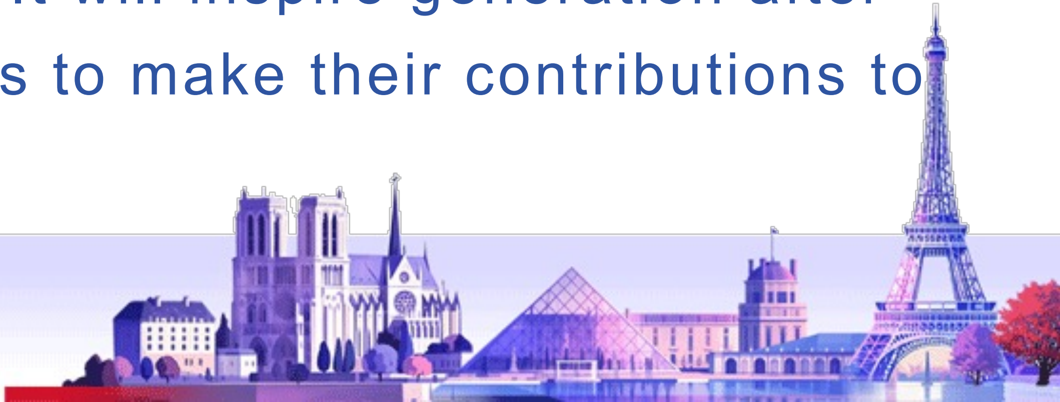
DeepSeek has the highest degree of openness

- DeepSeek supports the **MIT license**, which is business-friendly (one of the most permissive open-source licenses). It opens up the model weights and most of the code。
- There are **no restrictions** on model distillation, synthetic data, etc.
- There are also a great many technical **details** disclosed in the **papers** and open-source code.
- Among the top 10 large language models globally, DeepSeek has the highest degree of openness.



The open source with longtermism is the cradle of innovation

- From a deeper perspective, DeepSeek has truly grasped the spirit of open source. It starts from the goal of innovation for all mankind and embraces open source from a **strategic** rather than a tactical standpoint.
- This value is transforming China's open-source ecosystem, shifting it from being "closed-source imitators" to "**open-source innovators**", and fundamentally promoting the growth of the entire AI ecosystem.
- The success of DeepSeek has ignited the purest and most original spark of innovation in China’s developer community. It will inspire generation after generation of Chinese and global developers to make their contributions to innovation on a global scale.



THANK YOU

WeChat Subscription

Email: lijz@csdn.net

Linkedin: https://www.linkedin.com/in/jianzhongli