alt.ai begins construction of large language models with trillions of parameters

Pursuing the world's best speed and cost performance by designing backward from use cases

alt Inc. begins construction of large language models with trillions of parameters: Pursuing the world's best speed and cost performance by designing backward from use cases

TOKYO, JAPAN, May 8, 2024 /EINPresswire.com/ -- alt Inc. (https://alt.ai/en/), the Japan-based developer and distributor of Personal Artificial Intelligence (P.A.I.®) and AI clone technology (head office: Minato-ku, Tokyo; CEO: Kazutaka Yonekura), is pleased to announce that we have started to build a large language model (LLM) with several trillion parameters.

As a company that has been engaged in research, development and operation of LLMs, including natural language processing, for about 10 years, we believe it is essential to design and build LLMs by working backward from actual use cases in business and everyday life.

In addition to a high number of parameters, we aim to achieve the world's highest level of speed, computational efficiency, and cost performance, which are important indicators in actual operation.

Through this development, we aim to provide our existing generative AI products to end-users as more cost-effective services, as well as to establish a pioneering generative AI end-use case from Japan that sets a global precedent.

Awareness of issues and countermeasures against LLM competition
In the development of LLMs, alt is pursuing both more complex expressive power and a high degree of customizability. Taking into account the perspective of both developers and end-users, we envision a model that has trillions of parameters and is more extensive than the Japanese language, in order to achieve ease-of-use that surpasses existing models such as GPT. In LLMs, it

is important to strike a balance between speed and cost—in addition to the number of parameters, amount of data, and amount of compute, which are the basis of scaling rules. We will continue to manage these issues appropriately while providing practical services for our users.

For more details, please see:
https://drive.google.com/file/d/1HOtpCLTE9KPJHU2qP_vsxzthqP4oz66x/view?usp=sharing

Our aim in the direction of LLMs: the unexpectedly overlooked importance of speed and cost
In LLM development, it is true that increasing the number of parameters in a model improves its accuracy and expressiveness, but this alone may not be practical. For example, a model that takes as long as 30 minutes per prompt would spoil the user experience and no one would actually use it, so real-time response speed is required. Also, no one will use a model that costs 10,000 yen per prompt.

To maintain a practical level of response speed as a service, it is not enough to simply increase the processing power of the LLM; optimization at the hardware level, for example, by adopting an LLM-specific chip, is necessary. In addition, it is also essential to devise software aspects, such as building a high-speed architecture as an API, developing a communication technology infrastructure, and improving availability using decentralized technology.

To overcome the network latency problem, it is also important to increase speed by utilizing processing not only in the cloud, but also on the edge (terminal side). This will enable the construction of systems that can respond immediately to user requests.

In short, in the competitive LLM landscape, it is important to balance speed and cost as well as model accuracy and expressiveness. These must be managed appropriately while providing a service that is practical for users.

The underlying LLM/GPU energy problem
In addition, LLM operations and the use of GPU resources consume vast amounts of power. This energy problem not only affects the environment, but also has a significant impact on operating costs. In particular, LLM training and inference require a significant amount of compute, and the GPU resources to support this have high energy requirements.

Considering geographic differences in electricity costs, it is economical to operate data centers in areas with relatively low electricity costs. Geographic distribution of resources is therefore an effective strategy to reduce energy costs. However, this comes with challenges, such as delays in data transfers and regulations in certain regions.

On the other hand, it is also important to distribute energy consumption by processing

everything on the edge (terminal side) rather than on the server side. Edge computing reduces the amount of data transfer and shortens response time, thus reducing the burden on the server side, which in turn reduces overall energy consumption. The evolution of edge devices to allow more advanced processing to be performed on the terminal side can also expand the potential of this approach.

Overall, an effective strategy to address the energy problem of LLM and GPU resources is to decentralize resources to regions with lower power costs and to distribute energy consumption by leveraging edge computing. This will reduce operating costs while also minimizing environmental impact.

Based on an awareness of the above-mentioned issues and the current situation, we will accelerate the following efforts to address quality, speed, and cost, which are important to optimize end-use cases:

□□Large-scale construction of learning data
□□Construction and automation of instruction data
□□Automation of prompt engineering
□□Accelerated research on lifelong learning and metacognition to improve existing models (real-time continuous learning)
□□Accelerated research on the reproduction of outputs comparable to large-scale models using lightweight models (knowledge distillation, etc.)
□□Improvement of inference efficiency by introducing quantization and other methods
□□Exploring metacognitive processes to facilitate lifelong learning for personal AI
□□Research and development of LLM-specific chips (LLM-optimized ICs)
□□Development of a RAG database for the Japanese language
□□Further development of distributed computing infrastructure

Through this initiative, alt will expand the potential of AI technology, strengthen its business implementation capabilities, and provide new value to society. We will also actively promote collaboration with partners that contribute to this initiative.

□About alt Inc.
Founded in November 2014, alt is a startup that "aims to free people from unproductive labor" by creating P.A.I.®□ (Personal Artificial Intelligence) and AI clones. We also develop and provide various AI products that utilize our abundant AI elemental technologies, including generative AI, a proprietary LLM, and speech recognition technologies. As of April 2024, alt has raised over 10 billion yen.
https://alt.ai/en

<Media Inquiries to:>
Misako Nishizawa
e-mail: press@alt.ai

<Alliance Inquiries to:>
We provide AI solutions and support regardless of genre, including IT, finance, construction, logistics, media,
manufacturing, retail, and service industries.
Please feel free to contact us.

Junki Komura (AI Solutions Business Department)
e-mail: gptsolutions@alt.ai

Misako Nishizawa
alt Inc.
+81 3-6455-4677
press@alt.ai
Visit us on social media:
Facebook
Twitter
LinkedIn
YouTube
Other
