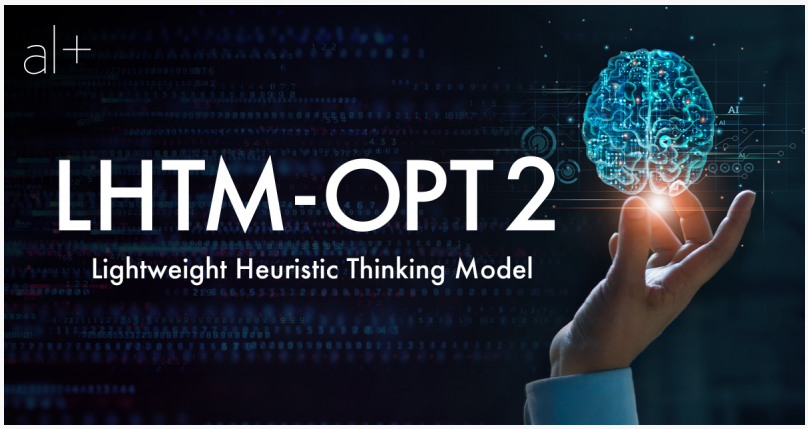
alt.ai’s LHTM-OPT2 achieves world's highest accuracy and inference speed as a lightweight LLM with Japanese RAG

Creating new use cases for Japanese AI with a lightweight language model and the language’s best inference capabilities



alt.ai’s LHTM-OPT2 achieves world's highest accuracy and inference speed as a lightweight LLM with Japanese RAG (Retrieval-augmented generation)Creating new use cases for Japanese AI with a lightweight language model and the language’s best inference cap

TOKYO, JAPAN, October 31, 2024 /EINPresswire.com/ -- alt Inc. (https://alt.ai/en/, head office: Minato-ku, Tokyo; CEO: Kazutaka Yonekura) is pleased to announce that its lightweight large language model LHTM-OPT2 has achieved the world's highest accuracy*1 for Japanese retrieval-augmented generation (RAG). LHTM-OPT2 is the latest version of our LHTM-OPT series of lightweight large language models, specially designed to optimize the accuracy of RAG.

*For reference: about LHTM-OPT: https://alt.ai/news_en/news_en-2398/
*For reference: alt Inc. releases world's first Japanese LLM on AWS Marketplace: https://alt.ai/news_en/news_en-2710/

Models in the LHTM-OPT series feature an optimized number of parameters practical for use on small GPU machines. The Japanese RAG accuracy of LHTM-OPT2 was evaluated using a dataset of RAG questions and answers from Wikipedia data (Wiki RAG dataset) developed independently by alt, plus the Japanese language subject dataset from the University of Tokyo entrance exam.

The Wiki RAG dataset was created by extracting specific paragraphs from Japanese Wikipedia, generating questions based on those paragraphs, and creating triples of [paragraph, question, correct answer]. This data was reviewed and corrected again by experts, resulting in high-quality RAG benchmarks.
For the second dataset, the prerequisite text (paragraphs) and questions for the Japanese

language subject questions of the University of Tokyo entrance examination*2 were used as input for the RAG, and experts evaluated the answers that the LLM generated from those paragraphs and questions.

LHTM-OPT2 achieved the same level of accuracy as GPT-4o (LHTM-OPT2: 91.0%, GPT-4o: 90.8%) on the Wikipedia RAG dataset. On RAG questions for the Japanese language section of the University of Tokyo entrance exam, LHTM-OPT2 achieved 94% of the accuracy of GPT-4o.

Furthermore, in the RAG evaluation, LHTM-OPT2 achieved higher accuracy than all other lightweight LLMs in Japan (LLMs with 10B or less parameters), and also recorded the highest scores for a lightweight LLM in the JGLUE (Japanese General Language Understanding Evaluation) benchmark and Japanese MT-Bench.*3

With the cooperation of SambaNova, we achieved an average speed of 500 TPS (tokens per second) and a maximum speed of 796 TPS in Japanese language inference. This speed is the highest ever recorded*4 for Japanese language LLM inference.

*1 World's highest accuracy and highest scores:
In an evaluation using the LLM/RAG benchmark, "RAG dataset using Japanese Wikipedia data developed by alt" achieved the top score in Japan among models with 10B or fewer parameters as the lightweight LLM.
(As of October 15, 2024. In-house research)

*2 Past University of Tokyo entrance exam questions and answers: https://www.u-tokyo.ac.jp/ja/admissions/undergraduate/e01_04.html

*3 Japanese MT-Bench is a benchmark test provided by Stability AI. In a performance evaluation on October 15, 2024, LHTM-OPT2 received the highest score of any lightweight LLM. A benchmark test is a method of measuring performance based on established standards, and Japanese MT-Bench uses GPT-4 as the evaluator.

*4 Record speed
According to ArtificialAnalysis.ai, among existing LLMs, Cerebras is the fastest at 2148 TPS, followed by SambaNova at the second fastest (462 TPS). However, alt and SambaNova are the first to achieve ultra-high-speed inference for an LLM dedicated to Japanese.
(As of October 15, 2024. In-house research)
https://artificialanalysis.ai/#providers

Through the development and provision of the LHTM-OPT series, alt will continue to provide more high-precision and efficient language models, aiming to become the "OpenAI of Asia" with its world-class technology. By providing our customers with the highest-quality solutions, we will promote efforts that contribute to improving the labor productivity of Japanese companies.

□For inquiries about LHTM-2/LHTM-OPT/GPT and other large language models solutions
□https://alt.ai/aiprojects/gpt/

□About alt Inc.
Founded in November 2014, alt is a company that "aims to free people from unproductive labor" by creating "P.A.I." (Personal Artificial Intelligence) and AI clones. In addition to AI GIJIROKU, a communication intelligence that utilizes speech recognition technology born from the development of an AI dialogue engine, we also develop and provide products, such as altBRAIN, AI Call Center, and CLONEdev, that provide solutions to various business issues through PoC (Proof of Concept).
https://alt.ai/en/

<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 (Business Headquarters)
e-mail: gptsolutions@alt.ai

<Media Inquiries to:>

Misako Nishizawa
alt Inc.
email us here
Visit us on social media:
Facebook
X
LinkedIn
YouTube
Other
