PRESS RELEASE
International Center for Quantum-field Measurement Systems for Studies of the Universe and Particles
WPI research center at KEK
World Premier International Research Center Initiative
KEK High Energy Accelerator Research Organization

May 16, 2025

International Center for Quantum-field Measurement Systems for Studies of the Universe and Particle (WPI-QUP) ,
Institute of Particle and Nuclear Studies (IPNS) ,
High Energy Accelerator Research Organization（KEK）

GPT-4o-Level Performance in Astrophysics & Cosmology with a Specialized 8-Billion-Parameter Large Language Model

Executive Summary

● Question

Can a highly specialized large-language model (LLM) built with only eight billion parameters match the performance of vastly larger and more expensive general-purpose AIs such as GPT-4o on cosmology tasks?

● Findings

Led by Prof. Tijmen de Haan of IPNS/WPI-QUP at KEK, the large language model “AstroSage-8B” was developed: an AI assistant trained on 250,000 paper preprints in astronomy, astrophysics, cosmology, and astronomical instrumentation. On the 4,425-question AstroMLab-1 benchmark the model gets 80.9% of the answers right, exceeding the performance of OpenAI’s GPT-4o while operating at roughly one thousandth the cost.

● Meaning

The study proves that given carefully curated data on a specific subject, the combination of continued pre-training, supervised fine-tuning, and model merging can yield a small, open-weight model that outperforms very large models. Not only does this lower the barrier for academic institutions with modest budgets to deploy powerful AI, it also paves the way for the development of autonomous research tools.

Overview

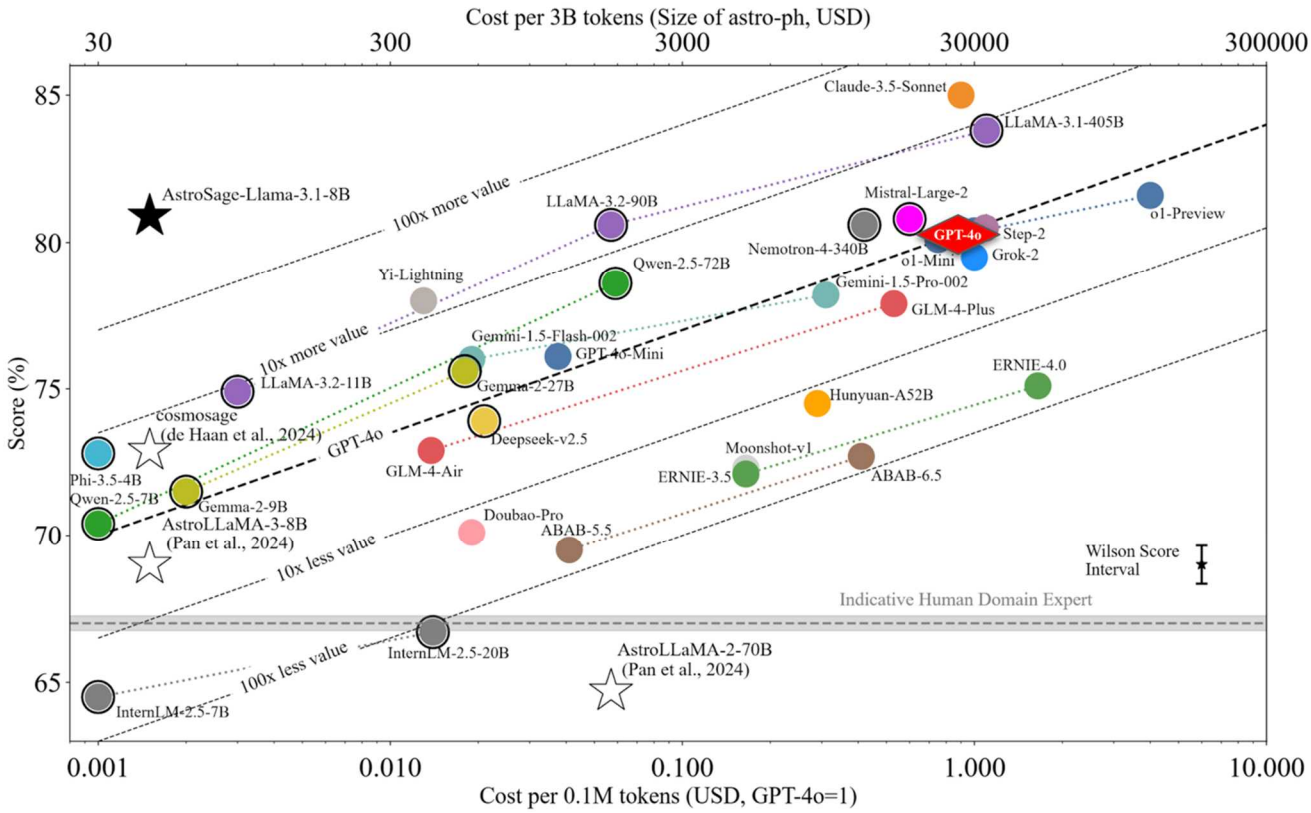
Can a smaller, highly specialized AI match or exceed the performance of giant, general-purpose AIs, and do so at a lower cost? That is the question Prof. Tijmen de Haan and his collaborators set out to answer. AstroSage-8B answers this question with a resounding “yes.” Most headline-grabbing AIs contain hundreds of billions to trillions of numerical “weights” and cost a fortune to train and operate. Working from KEK’s QUP and IPNS institutes, de Haan taught an 8-billion-parameter model—less than one hundredth the size of GPT-4o—to understand and reason in the domains of astronomy, astrophysics, astroparticle physics, cosmology, space science, and astronomical instrumentation (hereafter shortened to “astronomy”). These findings, demonstrating the power of specialized AI, were published on Apr 21st, 2025 in *Scientific Reports*.

The process to create AstroSage-8B was straightforward in principle, but demanding in practice. Following the method Prof. de Haan developed in [1], almost every paper in astronomy and cosmology published since 2007—about 250,000 documents—was gathered and converted into a machine-readable format. Prof. de Haan then trained the model on the Frontier exascale supercomputer, imbuing it with astronomical knowledge. He then taught the model how to act as a chatbot: to answer user queries with correct, helpful and clear answers. Unlike the rule-based AI systems of the 1990s, this teaching was done by example: millions of synthetic question–answer pairs were generated by LLM, then checked for quality using a different LLM. The 8.8 million Q&A pairs that were deemed of the highest quality were then used to fine-tune the astronomy-specialized model. Finally, the resulting weights were averaged with a general-purpose language model to instill capabilities not covered by the Q&A examples, such as the ability to have multi-turn conversations and to answer questions outside the astronomy domain.

The three-stage specialized training approach proved highly effective. On a 4,425-question expert benchmark [2], AstroSage-8B answers with 80.9 % accuracy—higher than OpenAI’s GPT-4o—while running roughly one thousand times cheaper. Because the model weights are openly licensed, any observatory, research organization, university, or high-school classroom can deploy a top-tier astronomy assistant on a single GPU. Researchers can use it to draft telescope proposals, debug data-analysis code, fill gaps in their knowledge, or brainstorm new ideas; students can quiz it for

clear explanations of redshift, exoplanet atmospheres, or the cosmic microwave background.

AstroSage-8B is a proof of concept: a compact, affordable AI assistant that rivals the largest commercial systems when focused on a well-defined scientific field.



Benchmark performance: AstroSage-8B (★) compared to other AI models on the expert AstroMLab benchmark [2]. The sloped lines show the approximate performance-cost tradeoff seen in most model families. AstroSage's performance slightly exceeds GPT-4o (♦) while operating at approximately 1000x lower cost, demonstrating that domain-specialized models yield simultaneously performant and cost-effective AI. Human expert performance is shown as the gray band.

Research group

This project was led by Tijmen de Haan—an assistant professor at KEK IPNS/QUP—in collaboration with an international team known as AstroMLab. This diverse group of researchers from all over the world consists of experts in astronomy, cosmology, natural language processing, and astrophysics data systems.

Future Work

This achievement raises the question: given that AstroSage-8B performs this well at just 8 billion parameters, what might be possible at a larger scale? Could a 10x larger model be the strongest LLM in the world within its domain of specialization? “We will know the answer to that soon,” said Tijmen de Haan, “Our next-generation model ‘AstroSage-70B’ is currently in training and so far, it looks very promising.”

Published paper

Title: AstroMLab 3: Achieving GPT-4o Level Performance in Astronomy with a Specialized 8B-Parameter Large Language Model

Authors: Tijmen de Haan, Yuan-Sen Ting, Tirthankar Ghosal, Tuan Dung Nguyen, Alberto Accomazzi, Azton Wells, Nesar Ramachandra, Rui Pan, Zechang Sun

Journal: Scientific Reports (Nature Portfolio)

https://www.nature.com/articles/s41598-025-97131-y

DOI: 10.1038/s41598-025-97131-y

References

[1] T. de Haan, “Cosmosage: A Natural-Language Assistant for Cosmology,” Astronomy & Computing 51 (2025) 100934.

[2] Y.-S. Ting, T. de Haan, et al., “AstroMLab 1: Who Wins Astronomy Jeopardy!?,” Astronomy & Computing 51 (2025) 100893.

Contact

International Center for Quantum-field Measurement Systems for Studies of the Universe and Particles (WPI-QUP) <qup_pr@ml.post.kek.jp>

Institute of Particle and Nuclear Studies (IPNS) <ipns-pr@ml.post.kek.jp>

High Energy Accelerator Research Organization (KEK) <press@kek.jp>