

Kimi K3 Tops New Benchmark of AI Models for Geological Reasoning

Groundtruth benchmark tests leading AI models on questions generated from real-world geological datasets

SINGAPORE, SINGAPORE, August 28, 2026 /EINPresswire.com/ -- Kimi K3 recorded the highest overall score on the new [Groundtruth benchmark](#), designed to test how well leading AI models can reason over novel real-world earth science information.

The Groundtruth benchmark tested models from OpenAI, Anthropic, Deepseek, Google, XAI and Moonshot, as well as the new stealth release OxAlpha (GLM 5.3 Flash), across three geological datasets spanning historical exploration records, technical commodity reports and academic geology. Rather than relying on a fixed bank of general-knowledge questions, Groundtruth generates new evaluations from geoscience source material, allowing models to be tested against unfamiliar datasets and region-specific data.

“

Rather than asking a standard set of questions, Groundtruth generates a fresh set from your own dataset, letting you compare models and agent harnesses on the data that matters to you.”

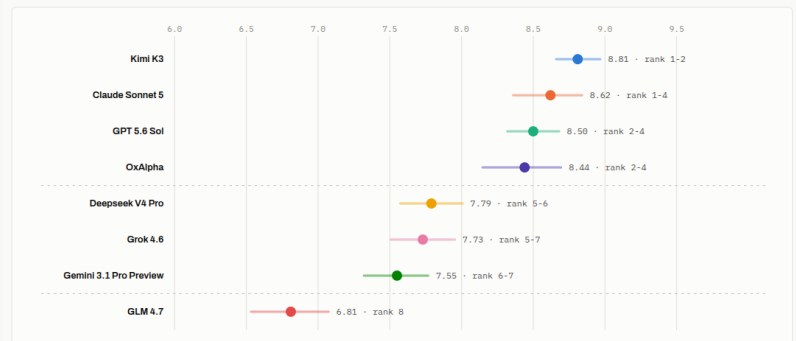
Jen Dodgson, Eigenform CEO

by Ox Alpha, which is currently available free.

The results highlight a growing problem in evaluating AI for specialist scientific work. General-purpose benchmarks can measure broad reasoning ability, but provide limited evidence of how

Tiers, not ranks

Mean score per question with a bootstrap 95% confidence interval. Where intervals overlap, the benchmark cannot order the models — and it says so. The models split into tiers, and only the boundary between tiers survives testing — the exact order inside a tier is not a claim.



Relative rankings of AI models on the Groundtruth benchmark

Kimi K3 achieved the highest aggregate score of 91%, closely followed by Claude and GPT instances. However, confidence intervals for the leading models overlapped, placing several systems in the same performance tier. The evaluation also found substantial differences in inference cost between models achieving comparable scores. Kimi, Claude, GPT and Deepseek offered broadly similar price-per-correct-answer rates, but were outcompeted on price

a model will perform when confronted with the reports, terminology, historical interpretations and incomplete evidence characteristic of a particular technical domain.

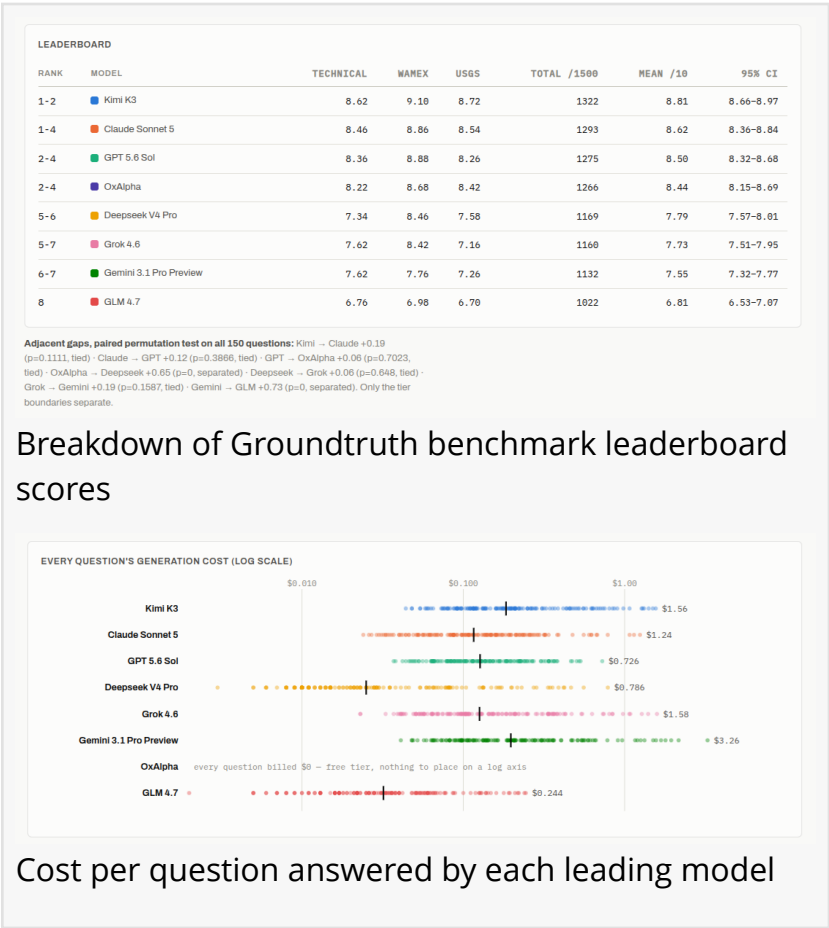
Groundtruth was developed by Singapore-based AI research company [Eigenform](#) AI, using their background in recursive self-improvement within complex dynamic environments to build a responsive alternative to existing benchmarks. The goal was to build a framework that would allow industry and academic users to create custom evaluations based on their own data to test alternative models, retrieval systems and harnesses.

Eigenform CEO Jennifer Dodgson said: “If you’re an exploration manager with 15 years of reports, logs, maps, assays and consultant studies sitting in SharePoint and a vendor wants you to buy their AI system they will tell you it scored 83% on some general academic benchmark, but that’s not important to you. You don’t want a model that knows bismuth can be a pathfinder for gold, you want a model that can work out whether it serves as a pathfinder in your specific terrain. That’s why we made the Groundtruth Dynamic Benchmark. Rather than asking a standard set of questions to test the sector knowledge of a model, Groundtruth generates a fresh set of questions and scoring rubrics from your own dataset, letting you compare models and agent harnesses on the data that matters to you.”

To create a representative leaderboard Eigenform worked with tenement mapping and marketplace service, NextMaps, and commodities analysis platform [MatchPoint](#) to create test datasets covering a variety of regions and stages in the exploration process.

MatchPoint CEO, Manessa Mungroo said: “AI is becoming an essential part of the resources value chain, and transparency is vital to its effective adoption. At MatchPoint, we have fully incorporated LLMs into our data extraction algorithms. Large volumes of unstructured data are now being processed into structured formats at scale, enabling discoveries and insights that would previously have taken days for an analyst to generate.”

Akira Raffael
Eigenform Pte Ltd



+65 9133 0611

alex@send.eigenform.ai

Visit us on social media:

[LinkedIn](#)

[X](#)

This press release can be viewed online at: <https://www.einpresswire.com/article/937552143>

EIN Presswire's priority is source transparency. We do not allow opaque clients, and our editors try to be careful about weeding out false and misleading content. As a user, if you see something we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.

© 1995-2026 Newsmatics Inc. All Right Reserved.