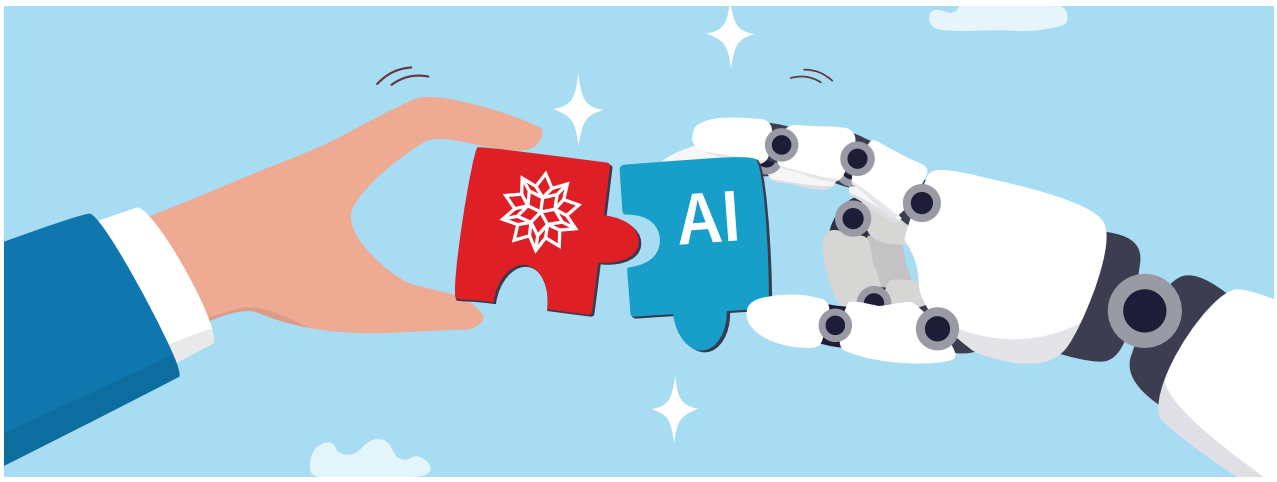
CUSTOMER SUCCESS STORIES

Quantifying Error Rates in Local LLMs with Wolfram Language

Industry: Artificial Intelligence **Applications:** Mathematics, Consulting



ABOUT

As artificial intelligence technology matures, organizations face more granular decisions about AI implementation. The challenge is increasingly moving from "What can AI do?" to "Which model performs best at a specific task?" Although larger models consistently outperform smaller ones, not every task requires the biggest available model; determining the sufficient model size for cost-effective performance remains an open problem. To help their clients navigate the problem, Novus-i2 used Wolfram Language to rigorously test the mathematical reasoning ability of local large language models. As a result, they developed a predictive framework that can recommend the optimal model size based on desired accuracy thresholds. This enables their clients to mitigate lengthy trial-and-error processes and reduce the time to find the right LLM fit, as well as accurately quantify the risk of failure.

HIGHLIGHTS

30–50%
reduction in infrastructure costs by selecting the optimal model for the task

13
local large language models evaluated, ranging from 1B to 72B parameters

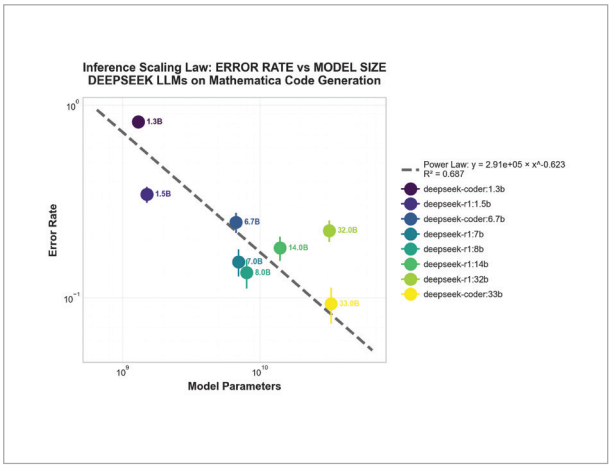
215
mathematical problems evaluated by each local model

THE CHALLANGE

Organizations deploying AI for mathematical tasks face a fundamental resource allocation challenge: determining the minimum model size required to achieve acceptable accuracy. Without clear performance thresholds, companies often default to larger, more expensive models than necessary or risk underperforming with insufficient computational power.

The rapid advancement of local models has made them a capable substitute for large, service-provided models, yet the sheer number of available options creates a new problem. Organizations must navigate dozens of models, each with different strengths and capabilities.

The selection challenge is compounded by the difficulty of rigorous evaluation. Unlike other AI applications where performance might be subjectively assessed, repartition reasoning requires objective verification—each answer is either mathematically correct or it isn't. This demands a computational platform capable of systematically evaluating mathematical accuracy, handling complex calculations and providing reliable performance comparisons across models.



Inference scaling visualization showing model parameters to error rate for the Deepseek model family.

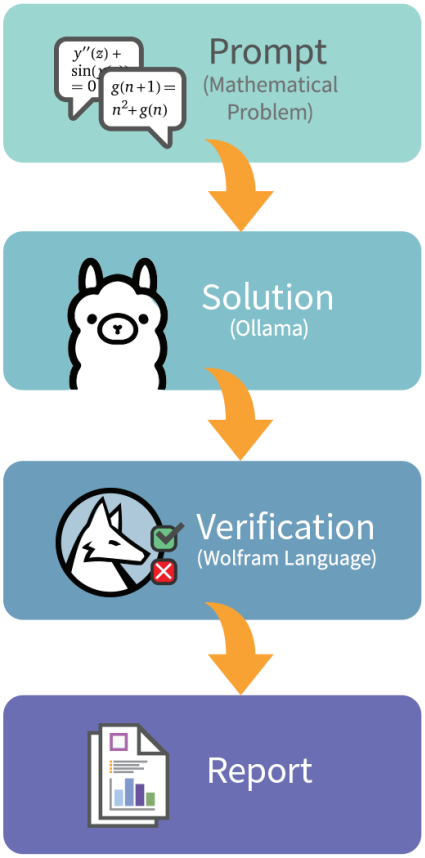
THE APPROACH

Novus-i2 developed a systematic testing framework using Wolfram Language to establish scaling laws for mathematical reasoning across different sizes of local language models. Wolfram Language emerged as the ideal platform for building this framework, combining comprehensive mathematical capabilities with native integration for local large language models.

The approach began with developing a comprehensive suite of mathematical problems spanning from simple algebra to complex symbolic mathematics. Each local model was then used to generate solutions, which fed into an automated pipeline that verified answers and produced detailed performance reports for each model. Having all components—problem generation, LLM interaction, mathematical validation and reporting—integrated within a single computational environment is invaluable for rapid iteration and consistent evaluation.

In a typical workflow, an LLM is provided with a mathematical problem and asked to develop a solution using Wolfram Language code. The solution is then passed to a Wolfram Language validator for an automatic assessment of its accuracy. Based on the result, the validator reports either a correct solution or an incorrect one. By tallying up all the correct solutions and all the incorrect solutions, Novus was able to get a more detailed sense of the model's performance compared to other models. Importantly, the framework for testing is fully automated and ready to be scaled up as new local models emerge in the constantly changing landscape of AI.

By systematically measuring error rates across model sizes within this standardized mathematical framework, Novus-i2 established predictive scaling laws that enable quantitative model selection decisions. With that, they are able to help organizations with critical deployment questions, such as “How much will task performance improve with a larger model?" instead of relying on trial-and-error approaches or subjective performance claims.



A simplified diagram of a single mathematical problem and a single LLM.

ACHIEVEMENTS

- **Established predictive scaling laws for mathematical reasoning**. Developed quantitative equations that predict model performance based on size, enabling data-driven selection decisions rather than trial-and-error approaches.
- **Delivered significant cost optimization for clients.** The framework helps organizations identify the smallest model that meets their accuracy requirements, eliminating overspending on unnecessarily large models.
- **Fully future-proof.** Due to the automated nature of Novus's framework, new local models can be added with minimal effort as they are released.

WOLFRAM

Best known for Mathematica and Wolfram|Alpha, Wolfram Research has been pioneering computational intelligence and scientific innovation for over three decades. Wolfram provides a highly integrated technology stack for multiparadigm data science, including the very latest methods in machine learning, computer vision, predictive analytics and automated reporting. All of these applications and many more are unified by Wolfram Language, a robust computational language with the largest integrated collection of algorithms ever assembled.

As well as providing technology, Wolfram's technical consulting group can lead or support your data projects from innovation to deployment.

TAKE YOUR PROJECT TO THE NEXT LEVEL

From data analytics to modeling, publishing APIs to developing neural nets, exploring new ideas to large-scale deployment…

Find out how the Wolfram System can transform your workflows and jump-start your projects.

+1-800-WOLFRAM (965-3726)
+1-217-398-7181 (outside US & Canada)
Wolfram Research, Inc.

+44-(0)1993-883400
(Europe & Middle East)
Wolfram Research Europe Ltd.

wolfram.com/customer-stories
wolfram.com/contact-us