How Sahara AI Powered Microsoft's Breakthrough in Multimodal AI Math Reasoning

When Microsoft Research needed to push the boundaries of AI evaluation, they turned to Sahara AI, and the results are reshaping how the world measures machine intelligence.

When Microsoft Research set out to push the boundaries of AI evaluation, they turned to Sahara AI. Together, we built the foundation of MATHVISTA—a state-of-the-art benchmark used to test leading models like GPT-4V, Bard, Claude, and Gemini. Powered by over 6,000 precision-labeled datapoints from Sahara AI, MATHVISTA is now reshaping how the world measures machine intelligence.

The next generation of AI won't be won by bigger models alone. It will be won by those who control the highest-quality, most specialized data, and that's where Sahara AI leads.

Traditional labeling services weren't equipped for the challenge. This project demanded reasoning-driven annotations, rigorous testing of annotators, and meticulous logical accuracy. That's why Microsoft Research chose Sahara AI, the premium platform for high-performance AI data services.

Raising the Bar: Where Others Struggled, Sahara AI Delivered

Data labeling today is no longer about simple tags or basic categorization. As AI models become more advanced, the need for complex, high-precision annotations continues to grow, and most labeling companies struggle to keep up.

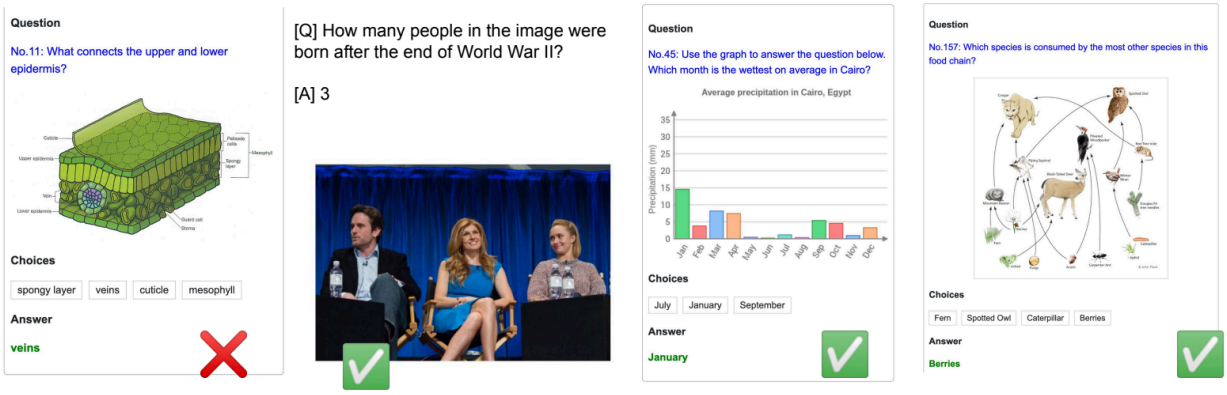
Building Microsoft's MATHVISTA proved just how high the bar has risen, demanding reasoning-driven annotations that most providers couldn't deliver, given the complexity and precision required:

> “This project posed significant challenges for other data labeling providers, particularly crowdsourcing platforms, as it required a deep understanding of complex instructions, rigorous testing of potential annotators, and meticulous labeling involving logical reasoning.”
>
> *— Hao Cheng, Principal Researcher, Microsoft Research*

To complete these data tasks, our annotators didn’t just “label” data; they performed cognitive work across multiple domains such as:

- **Arithmetic & Algebra** - Complex equation solving
- **Geometry & Statistics** - Visual pattern interpretation
- **Advanced STEM Logic** - Multi-step reasoning chains
- **Temporal Numeric Reasoning** - Time-series mathematical analysis
- **Numeric Commonsense** - Temporal knowledge

Each task demanded distinguishing between Deep Mathematical Reasoning (solving equations, interpreting graphs, algebraic structure) and Surface Recognition (counting, reading numbers, basic pattern matching).



Example of the types of images annotators had to review to determine if it involved mathematical reasoning or not.

Why Microsoft Chose Sahara AI

AI is entering a new phase where raw model size alone won’t cut it. Real, competitive intelligence depends on better data: more structured, more complex, and more bespoke.

That's what Sahara AI delivers: Not crowdsourced filler or vague approximations, but deeply logical, high-precision, enterprise-grade data that top-tier AI labs can actually trust.

For Microsoft, Sahara AI stood out in a competitive pilot phase by demonstrating:

- **Expert annotator selection** based on reasoning ability
- **Custom training modules** aligned with Microsoft's task requirements
- **Multi-phase quality assurance**, including reviewer oversight and consensus validation
- **Fast turnaround times** without sacrificing accuracy

Sahara AI labeled over 6,000 multimodal examples used in what became Microsoft's open-source MATHVISTA—a leading benchmark designed to stress-test models like GPT-4V, Bard, Claude, and Gemini on real-world math reasoning problems grounded in images, graphs, and text.

Since release, **MATHVISTA has become a trusted benchmark** for labs and researchers worldwide, with applications in testing and refining large multimodal models across academia and industry:

- **13K+ downloads** of the MATHVISTA dataset in the past month, with **275,864 downloads in all time.**
- Used in a peer-reviewed study evaluating **12 foundation models.** Results demonstrated that the best-performing model (GPT-4V) was only able to achieve an overall accuracy of 49.9% in multimodal math reasoning—10.4% below human performance.

This success underscores a larger truth: the future of enterprise AI depends on access to specialized, precision-labeled data, and Sahara AI is where leading institutions go when quality, speed, and trust can't be compromised.

In a space crowded with unproven claims, Sahara AI stands apart with real enterprise adoption and measurable impact. Microsoft Research, MIT, Amazon, and other global leaders already rely on our data services, underscoring Sahara AI's role in building the lasting infrastructure AI innovation depends on.

> "At Sahara AI, we believe the future of AI will be defined not by hype, but by proven results and lasting infrastructure. Our collaboration with Microsoft Research on MATHVISTA is a clear example of how specialized, high-quality data can set new standards for intelligence. This is only the beginning—we're committed to working with world-class partners to build the trusted data and infrastructure that enterprise AI truly needs."
>
> *— Sean Ren, Co-founder and CEO, Sahara Labs*

Our work on MATHVISTA is only the first step in our partnership with Microsoft. Both Microsoft Research and Sahara AI are enthusiastic about future collaborations, united in our commitment to shaping the next wave of AI innovation and setting new standards for what's possible.

Work With Sahara AI

From powering Microsoft Research's MATHVISTA to supporting leading AI labs worldwide, Sahara AI has built one of the most advanced data services platforms of its kind.

With global scale, multi-modality coverage, and a hybrid AI + human-in-the-loop approach, Sahara AI delivers the precision and reliability that modern AI development demands:

- **Global Reach** — Access to 200,000+ pre-vetted labelers across 35+ of countries, covering 45+ languages and dialects.
- **Multi-Modality Coverage** — Comprehensive support for text, images, video, and audio annotation.
- **Diverse Domain Expertise** — From complex mathematical reasoning to natural language understanding, finance, technology, and more.
- **AI + Human Synergy** — A combined AI-centered and human-in-the-loop labeling approach to ensure both speed and accuracy.

That's why enterprises like Microsoft, Amazon, Snap, and MIT trust Sahara AI when accuracy, speed, and dependability are non-negotiable.

About Sahara AI: Sahara AI is the agentic AI company dedicated to making AI more accessible and equitable. We build the core protocols, infrastructure, and applications that let personal agents anticipate and execute on your behalf. For this to work, infrastructure has to be trustworthy: verifiable execution, enforceable usage policies, and automatic value distribution across every tool, model, and service an agent touches. Sahara is building a growing suite of agent-powered applications on top of this foundation, including Sorin, your personal agent for global digital markets. Our solutions currently power AI agents and high-quality data for consumers, Fortune 500 enterprises, and leading research labs, including Microsoft, Amazon, MIT, Motherson, and Snap.

X | Discord | Telegram | LinkedIn