Scientists create the hardest AI test ever — and the results are surprising.

Humanity's Last Exam, consisting of 2,500 questions, shows that AI still has a large gap to bridge compared to human intelligence.

As AI systems increasingly score very highly on traditional academic assessments, researchers are beginning to realize a new problem: **tests once considered difficult are no longer challenging enough** .

Well-known benchmarks like Massive Multitask Language Understanding (MMLU), once considered rigorous standards, no longer accurately reflect the capabilities of modern AI models.

To address this problem, a team of nearly 1,000 researchers worldwide, including a professor from Texas A&M University, developed an entirely new test. Their goal was to create a test that was broad enough, difficult enough, and based on in-depth knowledge that current AI still struggles to handle.

The result is the Humanity's Last Exam (HLE) — a test consisting of 2,500 questions ranging from mathematics, natural sciences, humanities, ancient languages ??to many specialized academic fields.

Details about the project were published in the journal Nature, and the test was also featured on lastexam.ai.

One of the people involved in developing the test was Dr. Tung Nguyen, a lecturer in Computer Science at Texas A&M. He contributed to writing and editing many of the questions in the test.

According to him, when AI scores highly on tests designed for humans, many people easily assume that AI has reached a level of understanding comparable to humans. However, the new test shows that intelligence is not just about pattern recognition, but also involves depth of knowledge, context, and expertise.

The test is designed to find weaknesses in AI.

The goal of the test wasn't to 'beat' AI, but to identify areas where AI still has limitations. Experts worldwide participated in writing and evaluating the questions. Each question had a clear answer and was designed to prevent AI from quickly searching the internet.

Some questions required translating ancient Palmyrene texts, others involved identifying bird anatomy or analyzing the pronunciation of ancient Hebrew. Notably, if any AI system answered a question correctly, that question would be removed from the final test. This ensured the test was always beyond the capabilities of current AI.

Initial test results show that even top-tier AI models are still struggling with this test.

1. GPT-4o reached approximately 2.7%
2. Claude Sonnet 3.5 gained 4.1%
3. OpenAI o1 reached approximately 8%.

The most powerful systems currently available, such as Gemini 3.1 Pro and Claude Opus 4.6, achieve accuracy levels of approximately 40% to 50%.

Why do we need a new AI testing framework?

According to the research team, high scores on older tests do not necessarily mean that AI has achieved true intelligence. These tests only measured the ability to complete specific tasks, not a deep understanding.

Dr. Tung Nguyen also contributed 73 questions out of a total of 2,500 — ranking second among the authors. He is also the most prolific question writer in the fields of mathematics and computer science.

He argued that without accurate assessment tools, policymakers and developers could misunderstand the true capabilities of AI.

Despite its dramatic title, Humanity's Last Exam doesn't aim to assert that humans are being replaced. On the contrary, the test shows that there are still many areas where human knowledge and expertise play a crucial role.

According to the researchers, the goal is not to compete with AI, but rather to understand the strengths and weaknesses of the technology, thereby building a safer and more reliable system.

A long-term AI evaluation standard

Humanity's Last Exam is designed as a long-term assessment standard. Some of the questions are publicly available, while the majority remain confidential to prevent the AI ??from memorizing the answers.

According to the research team, despite the rapid development of AI, the gap between artificial intelligence and human intelligence remains quite large. Dr. Tung Nguyen emphasized that the scale of the project is what is most remarkable. It involves not only computer scientists, but also historians, physicists, linguists, and medical experts.

This very diversity has helped to highlight the gaps in current AI capabilities — and in a way, this is still the result of **human collaboration** .