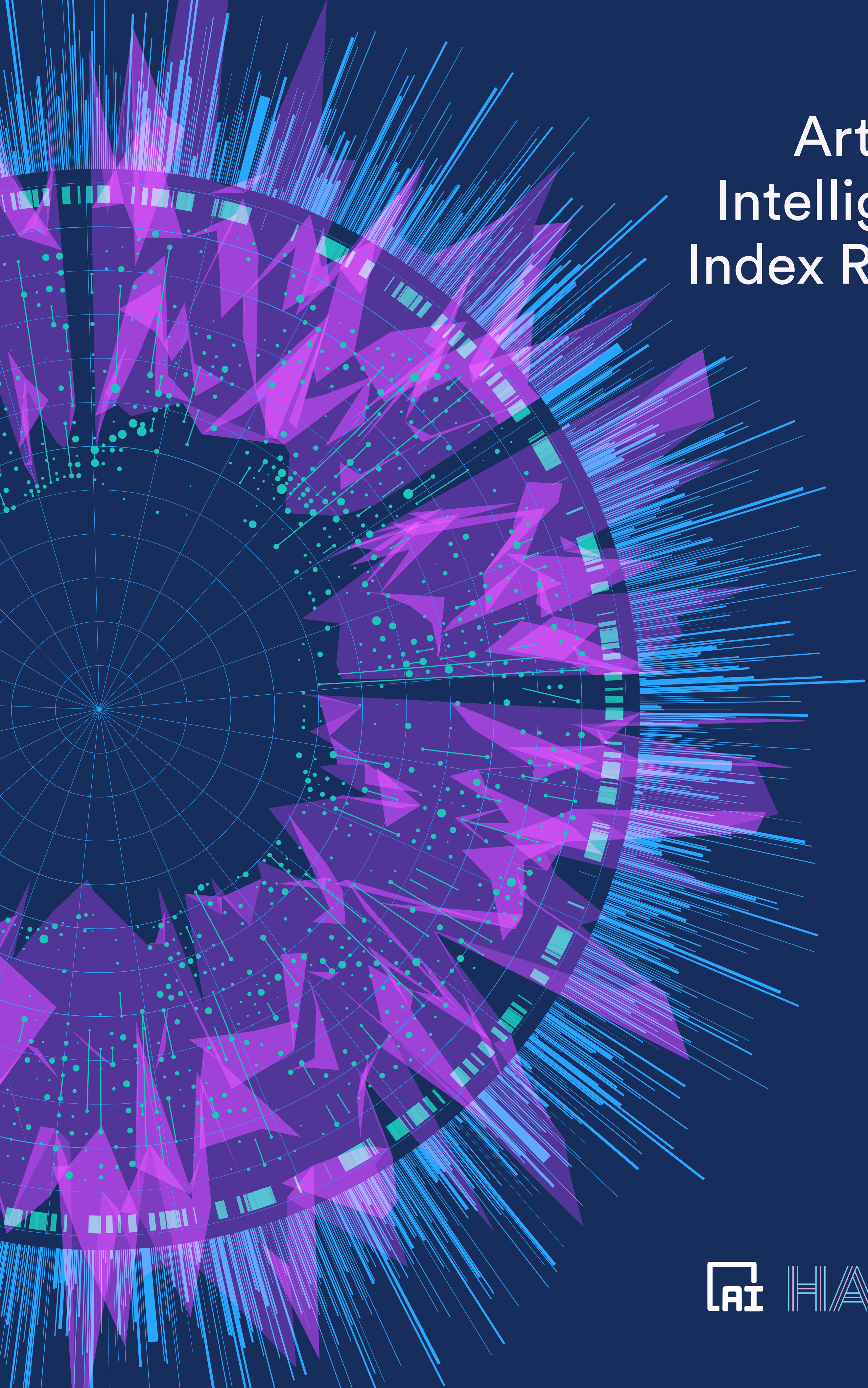

Artificial Intelligence Index Report 2024

Stanford University
Human-Centered
Artificial Intelligence

Message From the Co-directors

A decade ago, the best AI systems in the world were unable to classify objects in images at a human level. AI struggled with language comprehension and could not solve math problems. Today, AI systems routinely exceed human performance on standard benchmarks.

Progress accelerated in 2023. New state-of-the-art systems like GPT-4, Gemini, and Claude 3 are impressively multimodal: They can generate fluent text in dozens of languages, process audio, and even explain memes. As AI has improved, it has increasingly forced its way into our lives. Companies are racing to build AI-based products, and AI is increasingly being used by the general public. But current AI technology still has significant problems. It cannot reliably deal with facts, perform complex reasoning, or explain its conclusions.

AI faces two interrelated futures. First, technology continues to improve and is increasingly used, having major consequences for productivity and employment. It can be put to both good and bad uses. In the second future, the adoption of AI is constrained by the limitations of the technology. Regardless of which future unfolds, governments are increasingly concerned. They are stepping in to encourage the upside, such as funding university R&D and incentivizing private investment. Governments are also aiming to manage the potential downsides, such as impacts on employment, privacy concerns, misinformation, and intellectual property rights.

As AI rapidly evolves, the AI Index aims to help the AI community, policymakers, business leaders, journalists, and the general public navigate this complex landscape. It provides ongoing, objective snapshots tracking several key areas: technical progress in AI capabilities, the community and investments driving AI development and deployment, public opinion on current and potential future impacts, and policy measures taken to stimulate AI innovation while managing its risks and challenges. By comprehensively monitoring the AI ecosystem, the Index serves as an important resource for understanding this transformative technological force.

On the technical front, this year's AI Index reports that the number of new large language models released worldwide in 2023 doubled over the previous year. Two-thirds were open-source, but the highest-performing models came from industry players with closed systems. Gemini Ultra became the first LLM to reach human-level performance on the Massive Multitask Language Understanding (MMLU) benchmark; performance on the benchmark has improved by 15 percentage points since last year. Additionally, GPT-4 achieved an impressive 0.96 mean win rate score on the comprehensive Holistic Evaluation of Language Models (HELM) benchmark, which includes MMLU among other evaluations.

Top 10 Takeaways

1. AI beats humans on some tasks, but not on all. AI has surpassed human performance on several benchmarks, including some in image classification, visual reasoning, and English understanding. Yet it trails behind on more complex tasks like competition-level mathematics, visual commonsense reasoning and planning.

2. Industry continues to dominate frontier AI research. In 2023, industry produced 51 notable machine learning models, while academia contributed only 15. There were also 21 notable models resulting from industry-academia collaborations in 2023, a new high.

3. Frontier models get way more expensive. According to AI Index estimates, the training costs of state-of-the-art AI models have reached unprecedented levels. For example, OpenAI's GPT-4 used an estimated $78 million worth of compute to train, while Google's Gemini Ultra cost $191 million for compute.

4. The United States leads China, the EU, and the U.K. as the leading source of top AI models. In 2023, 61 notable AI models originated from U.S.-based institutions, far outpacing the European Union's 21 and China's 15.

5. Robust and standardized evaluations for LLM responsibility are seriously lacking. New research from the AI Index reveals a significant lack of standardization in responsible AI reporting. Leading developers, including OpenAI, Google, and Anthropic, primarily test their models against different responsible AI benchmarks. This practice complicates efforts to systematically compare the risks and limitations of top AI models.

6. Generative AI investment skyrockets. Despite a decline in overall AI private investment last year, funding for generative AI surged, nearly octupling from 2022 to reach $25.2 billion. Major players in the generative AI space, including OpenAI, Anthropic, Hugging Face, and Inflection, reported substantial fundraising rounds.

7. The data is in: AI makes workers more productive and leads to higher quality work. In 2023, several studies assessed AI's impact on labor, suggesting that AI enables workers to complete tasks more quickly and to improve the quality of their output. These studies also demonstrated AI's potential to bridge the skill gap between low- and high-skilled workers. Still, other studies caution that using AI without proper oversight can lead to diminished performance.

Top 10 Takeaways (cont'd)

8. Scientific progress accelerates even further, thanks to AI. In 2022, AI began to advance scientific discovery. 2023, however, saw the launch of even more significant science-related AI applications—from AlphaDev, which makes algorithmic sorting more efficient, to GNoME, which facilitates the process of materials discovery.

9. The number of AI regulations in the United States sharply increases. The number of AI-related regulations in the U.S. has risen significantly in the past year and over the last five years. In 2023, there were 25 AI-related regulations, up from just one in 2016. Last year alone, the total number of AI-related regulations grew by 56.3%.

10. People across the globe are more cognizant of AI's potential impact—and more nervous. A survey from Ipsos shows that, over the last year, the proportion of those who think AI will dramatically affect their lives in the next three to five years has increased from 60% to 66%. Moreover, 52% express nervousness toward AI products and services, marking a 13 percentage point rise from 2022. In America, Pew data suggests that 52% of Americans report feeling more concerned than excited about AI, rising from 37% in 2022.

Report Highlights

Chapter 3: Responsible AI

1. Robust and standardized evaluations for LLM responsibility are seriously lacking. New research from the AI Index reveals a significant lack of standardization in responsible AI reporting. Leading developers, including OpenAI, Google, and Anthropic, primarily test their models against different responsible AI benchmarks. This practice complicates efforts to systematically compare the risks and limitations of top AI models.

2. Political deepfakes are easy to generate and difficult to detect. Political deepfakes are already affecting elections across the world, with recent research suggesting that existing AI deepfake methods perform with varying levels of accuracy. In addition, new projects like CounterCloud demonstrate how easily AI can create and disseminate fake content.

3. Researchers discover more complex vulnerabilities in LLMs. Previously, most efforts to red team AI models focused on testing adversarial prompts that intuitively made sense to humans. This year, researchers found less obvious strategies to get LLMs to exhibit harmful behavior, like asking the models to infinitely repeat random words.

4. Risks from AI are becoming a concern for businesses across the globe. A global survey on responsible AI highlights that companies' top AI-related concerns include privacy, data security, and reliability. The survey shows that organizations are beginning to take steps to mitigate these risks. Globally, however, most companies have so far only mitigated a small portion of these risks.

5. LLMs can output copyrighted material. Multiple researchers have shown that the generative outputs of popular LLMs may contain copyrighted material, such as excerpts from The New York Times or scenes from movies. Whether such output constitutes copyright violations is becoming a central legal question.

6. AI developers score low on transparency, with consequences for research. The newly introduced Foundation Model Transparency Index shows that AI developers lack transparency, especially regarding the disclosure of training data and methodologies. This lack of openness hinders efforts to further understand the robustness and safety of AI systems.