

The lightning onset of AI—what suddenly changed? An Ars Frontiers 2023 recap

Enlarge / On May 22, Benj Edwards (left) moderated a panel featuring Paige Bailey (center), Haiyan Zhang (right) for the Ars Frontiers 2023 session titled, “The Lightning Onset of AI – What Suddenly Changed?”

Ars Technica

reader comments

1 with

On Monday, Ars Technica hosted our Ars Frontiers virtual conference. In our fifth panel, we covered “The Lightning Onset of AI—What Suddenly Changed?” The panel featured a conversation with Paige Bailey, lead product manager for Generative Models at Google DeepMind, and Haiyan Zhang, general manager of Gaming AI at Xbox, moderated by Ars Technica’s AI reporter, Benj Edwards.

The panel originally streamed live, and you can now watch a recording of the entire event on YouTube. The “Lightning AI” part introduction begins at the 2:26:05 mark in the broadcast.

[embedded content]

Ars Frontiers 2023 livestream recording.

With “AI” being a nebulous term, meaning different things in different contexts, we began the discussion by considering the definition of AI and what it means to the panelists. Bailey said, “I like to think of AI as helping derive patterns from data and use it to predict insights ... it’s not anything more than just deriving insights from data and using it to make

predictions and to make even more useful information.”

Zhang agreed, but from a video game angle, she also views AI as an evolving creative force. To her, AI is not just about analyzing, pattern-finding, and classifying data; it is also developing capabilities in creative language, image generation, and coding. Zhang believes this transformative power of AI can elevate and inspire human creativity, especially in video games, which she considers the apex of artistic expression.

Next, we dove into the main question of the panel: What has changed that’s led to this new era of AI? Is it all just hype, perhaps based on the high visibility of ChatGPT, or have there been some major tech breakthroughs that brought us this new wave?

Advertisement

Enlarge / Paige Bailey of Google during her Ars Frontiers 2023 panel on AI.

Ars Technica

Zhang pointed to the developments in AI techniques and the vast amounts of data now available for training: “We’ve seen breakthroughs in the model architecture for transformer

models, as well as the recursive autoencoder models, and also the availability of large sets of data to then train these models and couple that with thirdly, the availability of hardware such as GPUs, MPUs to be able to really take the models to take the data and to be able to train them in new capabilities of compute.”

Bailey echoed these sentiments, adding a notable mention of open-source contributions, “We also have this vibrant community of open source tinkerers that are open sourcing models, models like LLaMA, fine-tuning them with very high-quality instruction tuning and RLHF datasets.”

When asked to elaborate on the significance of open source collaborations in accelerating AI advancements, Bailey mentioned the widespread use of open-source training models like PyTorch, Jax, and TensorFlow. She also affirmed the importance of sharing best practices, stating, “I certainly do think that this machine learning community is only in existence because people are sharing their ideas, their insights, and their code.”

When asked about Google’s plans for open source models, Bailey pointed to existing Google Research resources on GitHub and emphasized their partnership with Hugging Face, an online AI community. “I don’t want to give away anything that might be coming down the pipe,” she said.

Generative AI on game consoles, AI risks

Enlarge / Haiyan Zhang of Microsoft during her Ars Frontiers 2023 panel on AI.

Ars Technica

As part of a conversation about advances in AI hardware, we asked Zhang how long it would be before generative AI models could run locally on consoles. She said she was excited about the prospect and noted that a dual cloud-client configuration may come first: “I do think it will be a combination of working on the AI to be inferencing in the cloud and working in collaboration with local inference for us to bring to life the best player experiences.”

Advertisement

Bailey pointed to the progress of shrinking Meta’s LLaMA language model to run on mobile devices, hinting that a similar path forward might open up the possibility of running AI models on game consoles as well: “I would love to have a hyper-personalized large language model running on a mobile device, or running on my own game console, that can perhaps make a boss that is particularly gnarly for me to beat, but that might be easier for somebody else to beat.”

To follow up, we asked if a generative AI model runs locally on a smartphone, will that cut Google out of the equation? “I

do think that there's probably space for a variety of options," said Bailey. "I think there should be options available for all of these things to coexist meaningfully."

In discussing the social risks from AI systems, such as misinformation and deepfakes, both panelists said their respective companies were committed to responsible and ethical AI use. "At Google, we care very deeply about making sure that the models that we produce are responsible and behave as ethically as possible. And we actually incorporate our responsible AI team from day zero, whenever we train models from curating our data, making sure that the right pre-training mix is created," Bailey explained.

Despite her earlier enthusiasm for open source and locally run AI models, Bailey mentioned that API-based AI models that only run in the cloud might be safer overall: "I do think that there is significant risk for models to be misused in the hands of people that might not necessarily understand or be mindful of the risk. And that's also part of the reason why sometimes it helps to prefer APIs as opposed to open source models."

Like Bailey, Zhang also discussed Microsoft's corporate approach to responsible AI, but she also remarked about gaming-specific ethics challenges, such as making sure that AI features are inclusive and accessible.

<https://arstechnica.com/?p=1941696>