

WORLD'S TOP AI MEDIA POWERHOUSE



AIM MEDIA HOUSE



AIM Research
Exclusive

**TOP DATA
SCIENCE
PROVIDERS**

Unpopular
Opinion

**SUBBARAO
KAMBHAMPATI**

OPENAI MAFIA

Nearly 75 Former OpenAI Employees
have Founded 30+ AI Startups.

SEPTEMBER 2024 EDITION

Your Only Destination for Everything AI & Beyond



Download Now!

• AI INSIDER •

34 | **Generative AI in Healthcare & Life Sciences in India**

Dr Rowland Illing completed his undergraduate paediatric training at St John's Medical College and Hospital, Bengaluru, almost two decades ago.





18 | Heeyo.ai is Here to Be Your Kid's Smart AI Buddy

Xiaoyin Qu is not your typical AI startup founder. The super-energetic entrepreneur recently unveiled a new learning app for kids called Heeyo AI.

22 | The Shakti-Man of India

The SHAKTI-man of India, Veezhinathan Kamakoti, the director of IIT Madras, has had a hectic few years. The genius behind India's first microprocessor has been trying really hard to build semiconductor capabilities in India, and put a stop on cheap imports from China.

26 | LLMs Evolve Backwards

SLMs or specialised models are going to be the future, alongside generalised models like GPT-4 or Claude 3.5 Sonnet.

30 | Why AI is Incredibly Smart and Shockingly Stupid

Yejin Choi, a leading computer scientist, is deeply obsessed with instilling true common sense in AI to prevent catastrophic failures, arguing that while AI's advancements are remarkable, building advanced models isn't for everyone—because "AI today is unbelievably intelligent and then shockingly stupid."

38 | Top Data Science Service Providers – PeMa Quadrant 2024

AIM Research's annual initiative to recognize leading Data Science Service Providers sets an industry standard, helping businesses quickly assess the potential of vendors.



42 | Why This Semiconductor Chip Company Chose India Over China

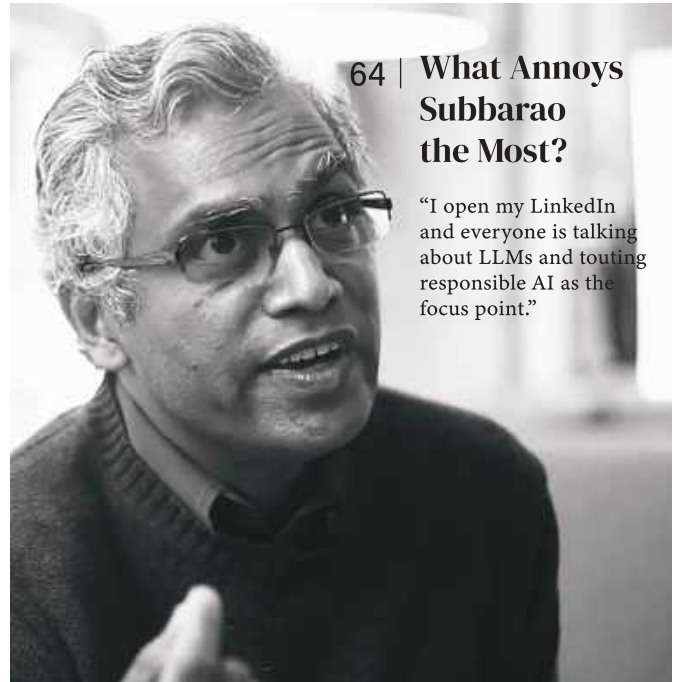
India will be an AI product focussed leader globally and a lot of innovation in AI will come from the country.

44 | The Future is AI Agents No Doubt.

Everybody is talking about them. In fact, no conversation in tech circles goes without the mention of 'AI agents' or 'agent AI' or 'compound AI systems' these days.

48 | The 'Linux Moment' in AI Has Finally Arrived

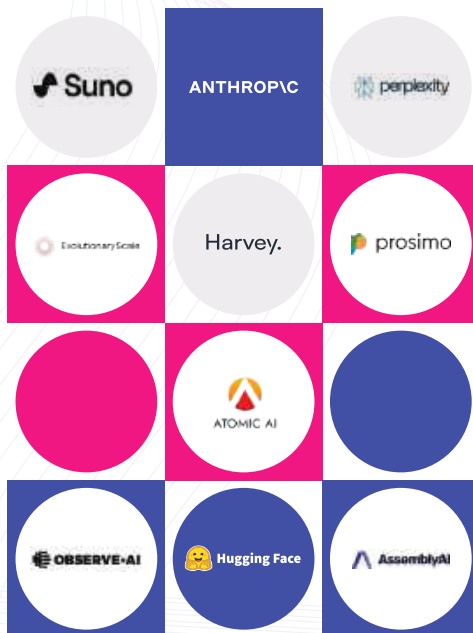
Developers can run inference on Llama 3.1 405B on their own infrastructure at roughly 50% of the cost of using closed models like GPT-4o.



64 | What Annoys Subbarao the Most?

"I open my LinkedIn and everyone is talking about LLMs and touting responsible AI as the focus point."

52 | Fastest Growing GenAI Startups



60 | AI Games Zone

Dive into mind-bending puzzles and challenges that spark your inner tech genius!

68 | If AI Works in India, it Can Work Anywhere

In the next 5-10 years, we would completely stop using the thumb, and use only voice-based interfaces," the Wadhvani AI CEO stated.

72 | Angels & Demons Of AI

In an exclusive interaction with AIM, Yann LeCun elaborated on his latest research and critiqued the chain of citations.



14 |



IS
OPENAI
“SAM”
of ALL
TRADES,
MASTER of NONE?



76 | The Machinecon USA 2024

78 | AWS Summit, Washington

79 | Databricks Data + AI Summit 2024

80 |



86 | Think Like Yoshua

In an exclusive interaction with AIM, Yoshua Bengio takes us through the three decade memory lane.

90 | Observability Tools Now Monitors LLMs

New Relic AI can now monitor AI-specific metrics like token usage, costs, response quality, as well as integration with their traditional application performance monitoring.

94 | Baidu Clocks 6,000 Driverless Rides Per Day in Wuhan, China

Baidu is clearly the darling of autonomous giants. A few months ago, Tesla chief Elon Musk visited China, the second largest market for the company, to release FSD (full-self driving) robotaxis in the region later this year.

94 | Integrating Generative AI and LLMs into Your Data

Integrating Generative AI and large language models (LLMs) into your data landscape represents a significant shift in harnessing artificial intelligence for data-driven solutions.

The story so far...



At the end of January last year, when we first wrote about the ‘OpenAI Mafia’ in our weekly AI newsletter, The Belamy, we took inspiration from a 2007 Fortune Magazine article titled ‘PayPal Mafia.’ Little did we know the impact our story would create in the AI ecosystem.

The story went viral in no time, and the rest is history. Coincidentally, our story was released a couple of months after the launch of the now-infamous ChatGPT on November 30, 2022.

Cut to the present. The so-called ‘OpenAI Mafia’ has grown significantly, with its numbers only increasing, giving birth to new-age AI startups like Anthropic, xAI, Pilot, Covariant, and Adept, among others.

Our latest investigation revealed that nearly 75 former OpenAI employees have founded around 30 AI startups. A notable recent departure is OpenAI co-founder Andrej Karpathy, who launched Eureka Labs to revolutionise education by combining generative AI with traditional learning methods. In contrast, Ilya Sutskever left OpenAI due to concerns about AGI safety and subsequently founded Safe Superintelligence.

“As we continue to track the remarkable journey of the OpenAI Mafia, we invite our readers to explore the world of AI innovations spearheaded by these visionary pioneers.”

Amit Raja Naik
Editor at AIM
AI Human



V24.09.01

FOUNDER

Bhasker Gupta

EDITORIAL

Amit Raja Naik, Mohit Pandey, Moushmi Sinha, Pritam Bordoloi, Vandana Nair, Sagar Sharma, Siddharth Jindal, S Vidyashree

RESEARCH

Anant Saxena, Abhijeet Adhikari, Navaneeth R, Yashashree Vivarekar

CREATIVE & DESIGN

Raghavendra Rao, Diksha Mishra, Raghu P, Chandni Yadav

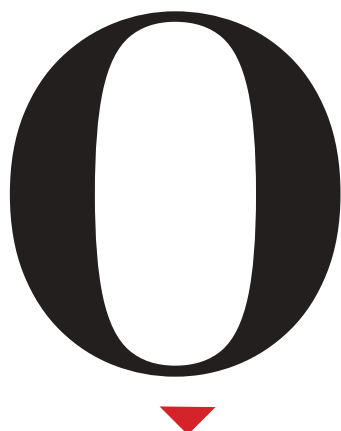
Founded in 2012, in *Namma Bengaluru*, AIM’s mission is to make technology a greater force for good by bringing about better-informed, more conscious technology decisions through authoritative, influential, and trustworthy journalism.

© 2024 AIM Media House LLC and/or its affiliates.
All rights reserved.

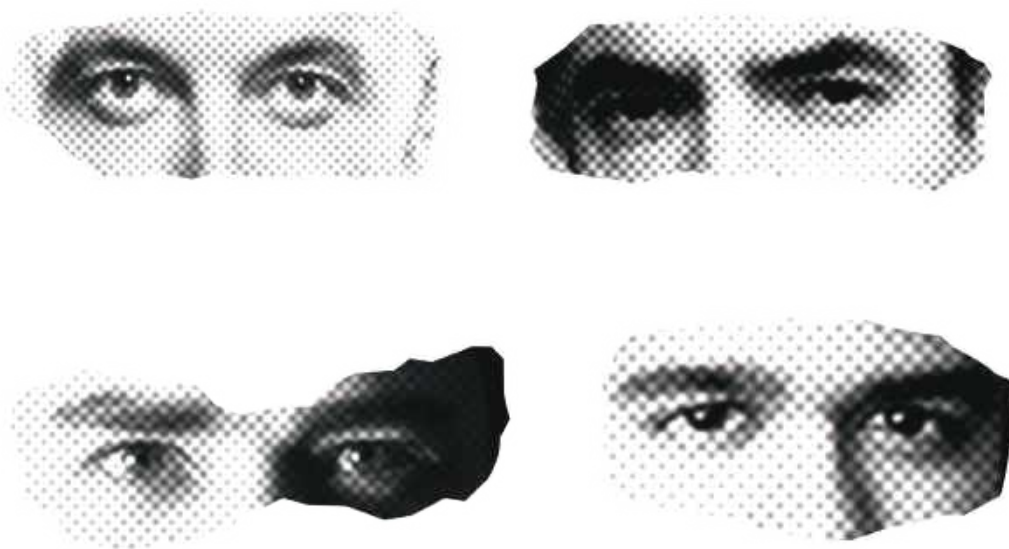
***For Internal Circulation Only



OPENAI MAFIA



OpenAI had approximately 2,500 employees as of last month and who knows how many more AI startups would spin out of the company. It is fascinating to see new tech entrepreneurs being born out of the OpenAI ecosystem, which is acting as a training ground for future AI leaders. “For a company like ours, the researchers and engineers that create the tech have far more impact than the CEO,” said Sam Altman. A quick look at its work culture shows that OpenAI encourages its employees to take creative risks in pursuit of advancements in AI and emphasis on collaboration, open communication and more.



“In 2015, Elon Musk, alongside Sam Altman and other prominent Silicon Valley leaders, including Peter Thiel, LinkedIn co-founder Reid Hoffman, and India’s former Infosys chief Vishal Sikka among others, pledged \$1 billion to set up OpenAI Inc, a non-profit AI research organisation, and the parent company of OpenAI LP, a ‘capped-profit’ company, institutionalised in 2019, to accelerate AI research and talent development.

The OpenAI mafia just got bigger. Last year, AIM set out to investigate how many former OpenAI employees had quit the company to launch their own ventures. The results are astonishing! Close to 75 employees have left OpenAI and founded around 30 AI startups.

Most recently, OpenAI co-founder Andrej Karpathy turned his decades-old passion for AI and education into a company called Eureka Labs. “It’s a new kind of school that is AI-native, combining generative AI with traditional learning methods,” said Karpathy.

However, Karpathy’s departure from OpenAI wasn’t as dramatic as Ilya Sutskever. “Nothing ‘happened,’” said Karpathy, “and it’s not a result of any particular event,

issue, or drama (but please keep the conspiracy theories coming as they are highly entertaining :)).”

Meanwhile, Sutskever left OpenAI, citing AGI safety concerns. Months later, he launched Safe Superintelligence. The company recently raised \$1 billion from NFDG, a16z, Sequoia, DST Global, and SV Angel.

Along with Sutskever, Jan Leike, OpenAI’s head of alignment, also resigned and joined Anthropic, an AI company founded by former OpenAI employees. At Anthropic, Leike announced he would continue working on the “super alignment mission”, focusing on scalable oversight, weak-to-strong generalisation, and automated alignment research. “I think the OpenAI mafia will reach #1 in a few years. Ex-employees of OpenAI in senior

positions left to start their own cutting-edge projects like Anthropic and xAI, and they are wildly successful,” posted a user on X. “Prediction: the group at OpenAI will make the PayPal mafia look like peanuts,” posted another user.

To be honest, OpenAI is an intriguing place to work. Recently, Steven Heidel, an employee at OpenAI, posted, “The only real downside to working at OpenAI is all the cool stuff you get to see but aren’t allowed to talk about yet.”

Super Dramatic Founders

Another notable startup emerging from OpenAI is Perplexity AI. Interestingly, when Aravind Srinivas joined OpenAI as a research intern, Elon Musk left the company. “I got into OpenAI for an internship in 2018, when Musk was still there,” recalled Srinivas in a recent interview. “I remember Musk left around the time I joined. I was idolizing this guy, and then he calls an all-hands meeting, announces that he’s no longer going to be involved, and swears at people left, right, and center before leaving the room. It was a lot of drama.”

Notably, last year Musk founded his own AI startup, xAI. One of xAI’s first major products is Grok, an AI chatbot integrated with X (formerly Twitter). Grok was unveiled in November 2023, with subsequent improvements leading to the Grok-2 model, which includes long-context capabilities and image understanding.

One cannot definitely overlook Anthropic, founded by ex-OpenAI employees Dario Amodei and Daniela Amodei, which is now giving OpenAI stiff competition. The company recently released Claude Sonnet, a model that rivals GPT-4o in performance.

Tim Salimans, a former research scientist and team lead at OpenAI, founded Aidence, a company specialising in generative modeling, semi-supervised and unsupervised deep learning, and reinforcement learning in 2015. Tim Shi, previously a member of technical staff at OpenAI, founded Cresta in 2017, which uses AI to help sales and service agents enhance their customer interactions.

In 2021, a former research intern at OpenAI, Anish Athalye, started Cleanlab, a data-centric AI platform that automatically finds and fixes errors in ML datasets. Another former research engineer at OpenAI, Taehoon Kim, recently founded Symbiote AI, a real-time, 3D avatar generation platform. Other prominent startups from the OpenAI alumni network include Pilot, Covariant, Adept, Living Carbon, Quill (now part of Twitter), and Daedalus, among others.

Will we see more? It’s highly likely that more employees might leave the company to start their own ventures. Last year, Altman bragged that OpenAI had a very lean team of about 375 employees. However, when he was fired, at least 745 of OpenAI’s approximately 750 staff signed a letter demanding his reinstatement, indicating that the company’s headcount is gradually increasing. [can](#)



“I remember Musk left around the time I joined. I was idolizing this guy, and then he calls an all-hands meeting, announces that he’s no longer going to be involved, and swears at people left, right, and center before leaving the room. It was a lot of drama.”



Maddie Hall
Special Projects



Petar Korponaić
Data Preparation Engineer



Stanislas Polu
Research Engineer



Andrej Karpathy
Co-founder



Elon Musk
Co-founder



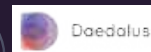
Shariq Hashme
Member Of Technical Staff



**Dario Amodei (VP of Research),
Daniela Amodei (VP of Safety & Policy)**



Matt Krisiloff
Founding Team Member



Jonas Schneider
Tech Lead



Ludwig Pettersson
Member of Technical Staff



Jeff Arnold
Head of Operations



Aravind Srinivas
Research Scientist



Peter Chen
Research Scientist



David Luan
VP of Engineering



Taehoon Kim
Research Engineer



Anish Athalye
Research Intern



Tim Shi
Member of
Technical Staff



Ilya Sutskever
Co-founder and
Chief Scientist



Tim Salimans
Research Scientist
& Team Lead





| By Siddharth Jindal

is
OPENAI
“SAM”
of
all
TRADES,
MASTER of NONE?



ChatGPT Plus, the premium version of OpenAI's popular chatbot, emerged as the primary revenue driver, contributing \$1.9 billion. The service boasts 7.7 million subscribers paying \$20 per month. Following closely is ChatGPT Enterprise, which brought in \$714 million from 1.2 million subscribers at \$50 per month. The API generated \$510 million, while ChatGPT Team added \$290 million from 80,000 subscribers paying \$25 monthly.

OpenAI is treading on thin ice. According to a recent report, the ChatGPT maker could lose as much as \$5 billion this year. Regarding costs, OpenAI is projected to spend nearly \$4 billion this year on renting Microsoft's servers to power ChatGPT and its underlying LLMs.

OpenAI's expenses for training, including data acquisition, could soar to nearly \$3 billion. Additionally, with a rapidly growing workforce of around 1,500 employees, its costs could reach approximately \$1.5 billion, driven partly by fierce competition for technical talent with companies like Google.

In total, OpenAI's operational expenses this year could reach up to \$8.5 billion. However, there is still hope. OpenAI recently generated \$3.4 billion annual revenue.

ChatGPT Plus, the premium version of OpenAI's popular chatbot, emerged as the primary revenue driver, contributing \$1.9 billion. The service boasts 7.7 million subscribers paying \$20 per month. Following closely is ChatGPT Enterprise, which brought in \$714 million from 1.2 million subscribers at \$50 per month. The API generated \$510 million, while ChatGPT Team added \$290 million from 80,000 subscribers paying \$25 monthly.

The question arises - "How long can OpenAI continue operating this way?" Without raising funds soon, the company may run out of cash to sustain its operations.

OpenAI may end up losing \$5 billion this year and run out of cash in 12 months unless it raises more money. Investors should ask: What is their moat? Unique tech? What is their route to profitability when Meta is giving away similar tech for free? Do they have a killer app? Will the tech ever be reliable? What is real and what is just a demo?" posted Gary Marcus on X. He may not be wrong. OpenAI has raised approximately \$13.3 billion in total funding to date from Microsoft.

OpenAI's Jugaad

OpenAI is cutting costs with GPT-4o Mini. At 15 cents per million input tokens and 60 cents per million output tokens, GPT-4o Mini is over 60% cheaper than GPT-3.5 Turbo. Recently, CEO Sam Altman posted that GPT-4o mini is "already processing more than 200B tokens per day!"

If we do some number crunching, processing 200 billion tokens daily, GPT-4o Mini costs OpenAI \$90,000 per day, compared to \$225,000 with GPT-3.5 Turbo. This transition allows OpenAI to save \$135,000 daily, showcasing enhanced efficiency and cost-effectiveness.

OpenAI Has no Money, no Moat

Regardless of these reports, OpenAI has now announced SearchGPT, a prototype with AI search features that delivers quick, timely answers, with clear and relevant sources. However, Perplexity AI, the much

“Guys, fine-tuned Llama 3.1 8B is completely cracked. Just ran it through our fine-tuning test suite and it blows GPT-4o mini out of the water on every task,” posted a user on X. “There has never been an open model this small, this good.”

touted Google-search alternative, already offers similar capabilities.

Moreover, Altman has also revealed that the voice feature on GPT-4o will be rolled out to Alpha users next week. However, Hume AI has already been leading the way with its Empathetic Voice Interface (EVI). Companies such as SenseTime and Kyutai are also developing voice-based products like SenseNova 5.0 and Moshi, respectively. At Google I/O 2024, Google introduced its AI agent Project Astra, which processes and integrates multiple modalities of data, including text, speech, images, and video. Apple has also upgraded Siri with ‘Apple Intelligence,’ enabling it to perform more actions for users.

From the perspective of LLMs, Meta released Llama 3.1, which performs even better than OpenAI’s GPT-4o and GPT-4o Mini in categories such as general knowledge, reasoning, reading comprehension, code generation, and multilingual capabilities. “Guys, fine-tuned Llama 3.1 8B is completely cracked. Just ran it through our fine-tuning test suite and it blows GPT-4o mini out of the water on every task,” posted a user on X. “There has never been an open model this small, this good.”

Within a day, Paris-based Mistral AI has also released Mistral Large 2, which offers substantial improvements in code generation, mathematics, and multilingual support, even outperforming Llama 3.1 in code generation and math abilities.

With a 128k context window and support for dozens of languages, including French, German, Spanish, and Chinese, Mistral Large 2 aims to cater to diverse linguistic needs. It also supports 80+ coding languages, such as Python, Java, and C++.

This puts the race to building LLMs and SLMs back into perspective. The case now seems that OpenAI, which was the pioneer in generative AI, is lagging behind when it comes to releases.

Not to forget, OpenAI has been postponing the launch of Sora, which is now expected later this year. Meanwhile, the video generation space has seen a surge of startups like Kling, RunwayML, and Luma AI. Recently, Kling AI, the Chinese company, announced the global launch of Kling AI International Version 1.0. Simply put, OpenAI is certainly a jack of all trades, master of none.

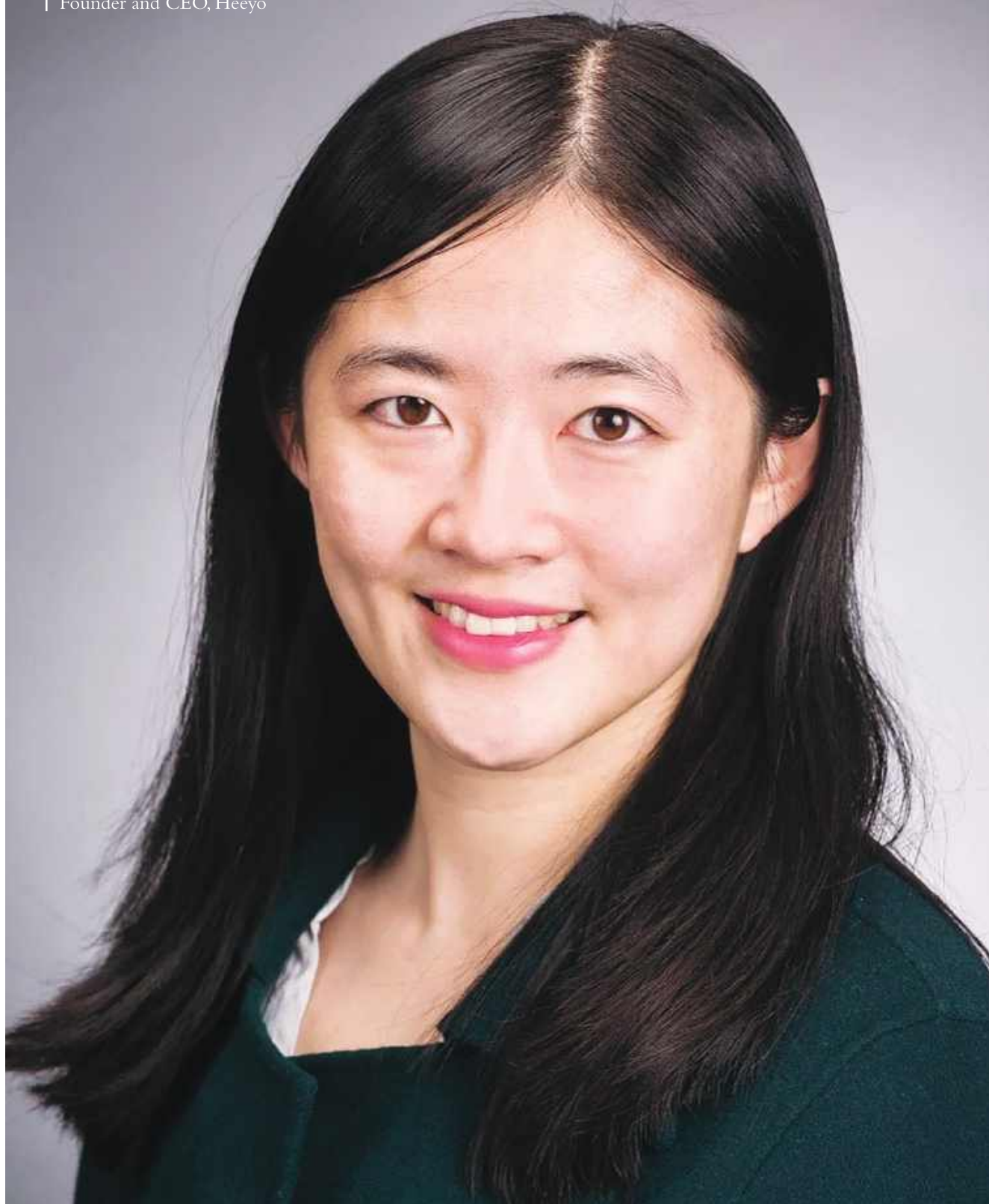
One thing OpenAI excels at is stealing the limelight from Google. Yesterday, Google DeepMind achieved a major breakthrough with its latest models, AlphaProof and AlphaGeometry 2, which solved four out of six problems from this year’s International Mathematical Olympiad (IMO), achieving a score equivalent to a silver medalist in the competition. “Advanced mathematical reasoning is a critical capability for modern AI. Today, we announce a major milestone in a longstanding grand challenge: our hybrid AI system has attained the equivalent of a silver medal at this year’s International Math Olympiad!” posted Google DeepMind chief Demis Hassabis, who himself won the London Mind Sports Olympiad for five consecutive years from 1998 to 2003.

However, a few hours later, OpenAI grabbed attention with its new Google alternative, SearchGPT. It combines OpenAI’s AI models with real-time web information to provide fast and relevant answers to user queries. Currently in the prototype phase, it is available to a limited group of 10,000 test users. 



Xiaoyin Qu

Founder and CEO, Heeyo



By Vandana Nair

Heeyo.ai is Here to Be Your Kid's Smart AI Buddy

19 | AI SPOTLIGHT

Xiaoyin Qu is not your typical AI startup founder. The super-energetic entrepreneur recently unveiled a new learning app for kids called Heeyo AI.

Dropping out of Stanford Graduate School of Business to start her entrepreneurial venture and eventually in the learning and teaching space, stands testament to her unwavering commitment to revolutionise education with AI.

During her late-night discussion with AIM, Qu enthusiastically explained the app, revealing that its vision stemmed from her own childhood desire for a coach who could guide her inventive and curious mind. "When you are 3 years old, you are just learning to talk. When you are 6 to 7, you can say multiple sentences. When you are 8 or 9, is when you can handle a lot of very open-ended creation. So, that's why AI actually needs to change the way they talk to you based on your cognitive abilities, and we have both systems that can cater to that," said Qu, founder and CEO of Heeyo AI, in an exclusive interaction with AIM.

Defining themselves as 'smart AI friends that help kids learn,' California-based Heeyo AI received a \$3.5M seed fund from some of the biggest tech players in the market including OpenAI startup fund, Amazon Alexa Fund, Par VC, Charge Ventures, StoryHouse Ventures and other top VCs. The startup is also funded by teams from top universities including Stanford and Harvard.

The learning platform provides over 2000+ learning games where the AI figure can speak in over 20 languages through fun avatars such as pandas and dragons.

Heeyo utilises various AI models, including text-to-speech, speech recognition, text-to-image, and text-to-music. They use specific models like OpenAI and Stable Diffusion for tasks such as content creation, translation, and image generation, with each project involving multiple models in a step-by-step process.

“

When you want to do a trivia on some Hindu traditions or Chinese traditions, you can actually build it up. In the US, we see some people are doing, like, Bible stuff. It's like Bible parenting.”



Social Emotional Learning

Heeyo is leveraging its AI platform, which is ideal for 3 to 11 year-olds, to address one of the biggest challenges of social emotional learning in children which is essentially helping kids improve their social skills. Qu highlighted the importance of teaching children how to make friends, express themselves, handle rejection, and respond appropriately when meeting someone new, and believes Heeyo will address all that. "That's a relatively new thing that all Silicon Valley parents want to be interested in," said Qu.

Furthermore, Heeyo allows anyone including parents, educators, and kids to design their own learning games, and make them based on different cultures. "When you want to do a trivia on some Hindu traditions or Chinese traditions, you can actually build it up. In the US, we see some people are doing, like, Bible stuff. It's like Bible parenting," said Qu.

Previous to this, Qu started Run The World, a leading virtual event platform, which she eventually sold last year. She has also led product management at Facebook and Instagram and led marketing & business development for Atlassian's Asian market.

AI for the Physical World

Qu is also ecstatic with the kind of funding Heeyo has received, and considers to be very lucky to have OpenAI invested in it. Not just OpenAI, but being backed by Amazon Alexa too, the potential for future products is plenty.

Qu highlighted how with many children already using Alexa, she believes that their content and interactive experiences could integrate well with Alexa to help even more kids. "In the long run, we plan to partner with toy companies, like family animals, and some of that. So, that's going to happen. But, right now, we're starting more from a content standpoint to make sure we have the right content," said Qu.

Interestingly, children's friendly AI robot Moxie, was also built with the intention of helping children develop social skills. Talking about it, Qu mentioned that Moxie is built for kids with special needs such as autism and has seen many adopt them. However, the price range of it makes it inaccessible for all. "We could imagine that Moxie costs \$899 and that's pretty much not affordable for most families," said Qu, who also said that Heeyo is accessible for all kids. The application is available for free and can be downloaded via the Apple or Google Play store.

With AI entering the education domain, a number of universities are already tying up with companies offering AI-related educational services. Recently, Andrej Karpathy, one of the former founders of OpenAI, launched Eureka Labs, which is an AI education company that aims to transform traditional teaching methods with generative AI.

Furthermore, with AI helping with automating a number of administrative and repetitive tasks, teachers are free'd up for dedicating time for personalised interactions with students, thereby enhancing the quality of education. With advanced interactive modes of teaching emerging through AI, learning and even teaching is continuously transforming for the better. Heeyo AI not only offers learning games but also allows parents, educators, and kids to design their own games, including those based on specific ethnic cultures and traditions. 



“

In the long run, we plan to partner with toy companies, like family animals, and some of that. So, that's going to happen. But, right now, we're starting more from a content standpoint to make sure we have the right content.”



**21-22
NOV**

**SANTA CLARA
CONVENTION CENTER,
CALIFORNIA**



CIM
PRESENTS

CYIPHERUSA

GENERATIVE AI
CONFERENCE FOR ENTERPRISE LEADERS
AND STARTUP FOUNDERS

Veezhinathan Kamakoti
Director, IIT Madras



THE SHAKTI-MAN OF INDIA

By Mohit Pandey



The SHAKTI-man of India, Veezhinathan Kamakoti, the director of IIT Madras, has had a hectic few years. The genius behind India's first microprocessor has been trying really hard to build semiconductor capabilities in India, and put a stop on cheap imports from China. "This is something we have achieved over the past 10 years; I have completely given up every other thing to work for SHAKTI," he told AIM in an exclusive interview.

Kamakoti started building the SHAKTI microprocessor in 2014, an open processor built on top of the open RISC-V architecture, which is funded by the Ministry of Electronics and Information Technology. He said that there is a need for building hardware that supports building AI models within the country, which led him to take the charge with the SHAKTI microprocessor.

For the past few years, Kamakoti has been rallying for an increase in the investments for semiconductor startups, which is still viewed as a risky area for many VCs. "I am only Saraswati, not Lakshmi for these startups," Kamakoti said, and informed that there are five companies coming out of the IIT Madras lab: Incore, Mindgrove Technologies, Chakra Electronics, Vyoma Systems, and SecurWeave. All of them are building on top of the SHAKTI core, but he is only the mentor for these companies. "It is very interesting to see that we have an indigenous processor ecosystem in India," he added.

Building trust for the chip

Since completing his first course in AI in 1992 from IIT Madras, Kamakoti has been finding ways to use AI for social good and solving complex problems within India. This led him to start focusing on hardware and eventually working on SHAKTI, which is also focusing on edge AI use cases.

Kamakoti believes that it is rather difficult to build trust in the semiconductor business because investors need to have full faith when investing in these startups. "Though the government is providing full support for improving the semiconductor industry in India, investors

are still hasty and cautious when it comes to investing, as it is a very costly affair, and cheaper alternatives are always available. We are now closing the gap," he added. "Today, Shakti C-class is very stable and we have made multiple chips with it. We are ready to give them to startups as well, which is not a joke," said Kamakoti highlighting that they are making it open for everyone for democratising processor based SoC (system of chips) design in the country. "Unlike a software company which has a cloud and is developing a software solution with demos, hardware startups take a lot more trust and investment, which the government is promoting, but we need a lot more VCs to take interest as well," Kamakoti emphasised.

BharatGPT can be the Copilot for India

Kamakoti was the chairperson of the AI Task Force, which was constituted in 2016 by the Ministry of Commerce and Industry. The AI Task Force report highlighted how AI can be used as an economic transformer for the country. Apart from this, currently he is also a crucial member of the BharatGPT initiative, which was started by IIT Bombay.

“ Though the government is providing full support for improving the semiconductor industry in India, investors are still hasty and cautious when it comes to investing, as it is a very costly affair, and cheaper alternatives are always available. We are now closing the gap.”

Highlighting specialised hardware for every field including IoT, network processing, and computing, Kamakoti said that it is also essential for LLMs to be built and catered for specific domains. He believes that this is where BharatGPT is going to play a crucial role. “The initiative requires validation from experts from every domain, and that is what BharatGPT stands for,” he explained.

Giving examples of training small models for the legal and medical sectors and how they can assist professionals in those field, Kamakoti said that the open source approach of BharatGPT will help it become the Copilot for different domains. “It is not a search engine, but a technology that is doing a lot of NLP, and giving you contextual output”, which he believes would help people approach every field with a lot more confidence.

Reflecting on how there are thousands, if not millions of documents in Indic languages, it is the need of the hour to extract knowledge from them, including historical facts and science. “For example, there is some proof that Newton’s laws were present in our ancient texts at least a 1000 years ago,” and this can benefit the whole of the world.

LLMs are like faithful kids

Talking about generalised models like ChatGPT and Bard, Kamakoti said that LLMs are like faithful kids. “If you give them relevant data, they



“

Unlike a software company which has a cloud and is developing a software solution with demos, hardware startups take a lot more trust and investment, which the government is promoting, but we need a lot more VCs to take interest as well,” Kamakoti emphasised.

will do exactly what you want them to.” He said that there is a dire need for data scientists and to focus on extracting and gathering information for tokenising these Indic models. “This would obviously require a lot of money and compute,” highlighting his love for farming, Kamakoti said that while we are also talking sustainability with AI models, they are also very necessary for helping various sectors such as agriculture. “There is a lot of knowledge hidden in rural India, which needs to be brought out,” he added, highlighting that every field, even within the same domain, requires different amounts of data.

Focusing on the recent funding of INR 110 crore from Sunil Wadhvani to IIT Madras for building a school for AI and data science, Kamakoti recalled that a reporter asked him if the same money could instead be spent on building 10-12 schools in India. To this, he replied, “I can use AI to enhance the capabilities of the teacher. I don’t want to replace the teacher, but augment them, in the end creating a very good Co-pilot.”

MachineHack

The Modern Network For Data Professionals

AI Hackathons | Assessments | Developer Advocacy

MachineHack is an online platform for machine learning hackathon and job opportunities. We help connect the machine learning engineers, data scientists and software engineers with recruiters by reducing the recruitment turnaround time using tailored assessments. Candidates are evaluated based on assessments defined by hiring managers.

200+

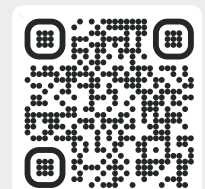
Tailored Online Hackathon

60,000+

Active AI Engineers

5000+

Strong ready-made question bank



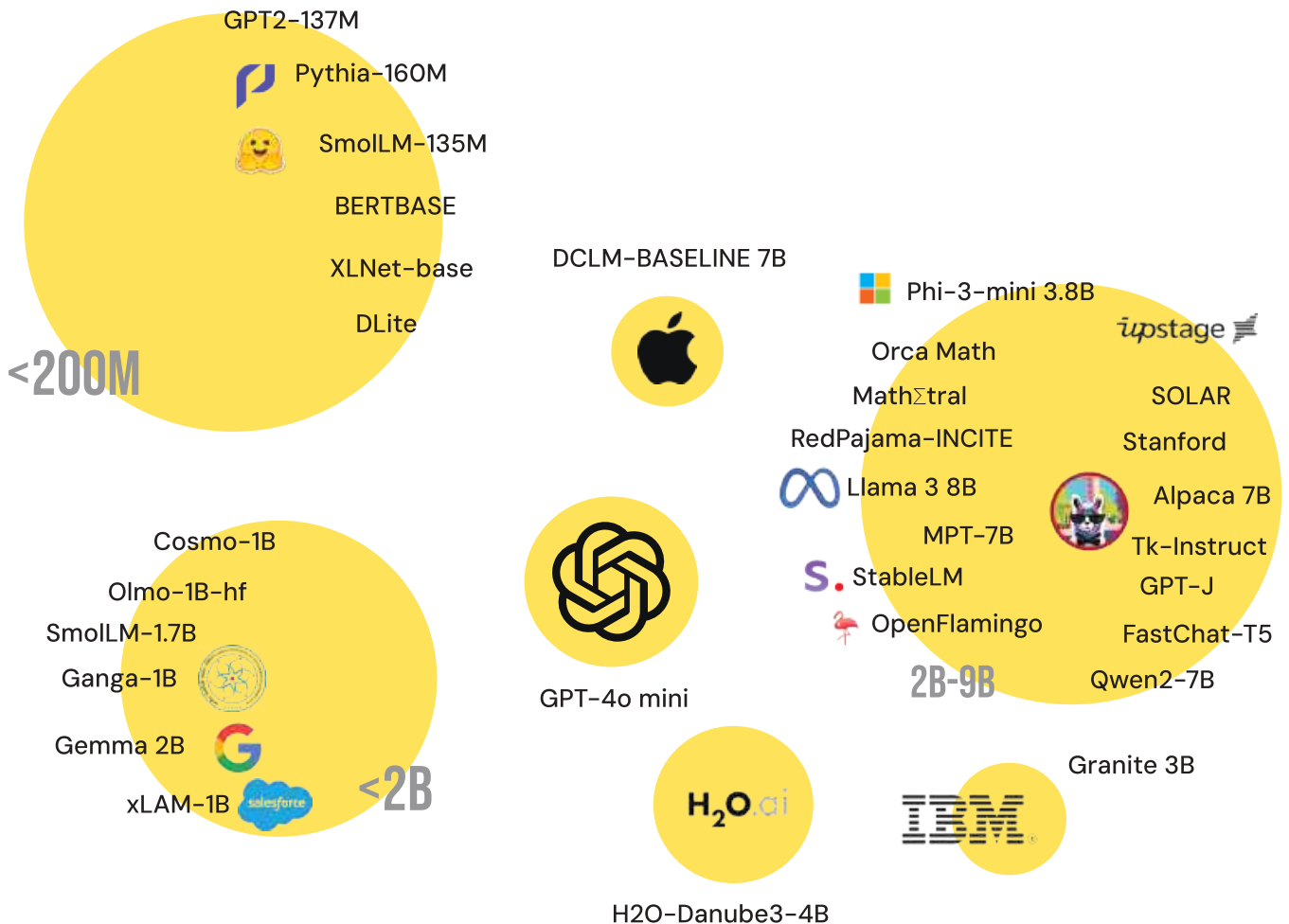
For any questions or clarifications!
info@aimmediahouse.com

LLMs Evolve

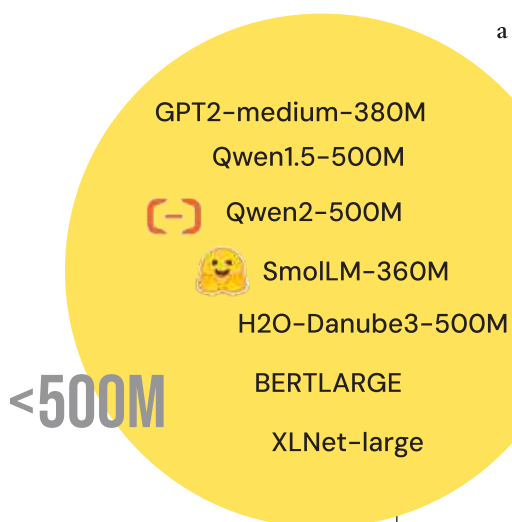
Backwards

SLMs or specialised models are going to be the future, alongside generalised models like GPT-4 or Claude 3.5 Sonnet.

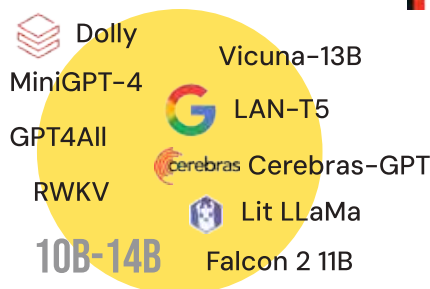
26 LLM MODEL SIZE COMPETITION IS INTENSIFYING, BACKWARDS



1-Bit LLM



Mistral NeMo 12B



The era of small language models has just begun. “LLM model size competition is intensifying... backwards!” quipped OpenAI cofounder Andrej Karpathy, reflecting on the recent announcements about GPT-4o mini and other small language models (SLMs).

Last few months saw a notable rise of several SLMs. OpenAI released GPT-4o mini, a highly cost-efficient model designed to expand AI applications by making intelligence more affordable. Priced at \$0.15 per million input tokens and \$0.6 per million output tokens, GPT-4o mini is 30x cheaper than GPT-4o and 60% cheaper than GPT-3.5 Turbo.

On the other hand, Paris-based AI startup Mistral AI, in collaboration with NVIDIA, released Mistral NeMo,

a 12 billion parameter model with a 128k token context length. Benchmarked against Gemma 2 9B and Llama 3 8B, it excels in reasoning, world knowledge, and coding accuracy. Mistral NeMo is particularly effective in English, French, German, Spanish, Italian, Portuguese, Chinese, Japanese, Korean, and Hindi. Not only that, Mistral also launched Math tral, a specialised 7B for advanced mathematical reasoning scientific exploration.

A few days ago, Hugging Face also released series of compact language models called SmoLLM, available in three sizes: 130M, 350M, and 1.7B parameters. These models are perfect for use on local devices such as laptops and phones, eliminating the need for cloud-based resources and significantly reducing energy consumption.

Similarly, H2O.ai introduced H2O-Danube3, a new series of SLMs to bring AI capabilities to mobile devices. The series includes two models: H2O-Danube3-4B, trained on 6 trillion tokens, and H2O-Danube3-500M, trained on 4 trillion tokens.

Apple too entered the game, releasing a 7B model called DCLM-BASELINE 7B, along with its weights, training code, and dataset. Trained on 2.5 trillion tokens from open datasets, the model primarily uses English data and features a 2048-token context window.

In China, Alibaba released Qwen’s Base and Instruct models in five sizes, including Qwen2-0.5B, Qwen2-1.5B, Qwen2-7B, Qwen2-57B-A14B, and Qwen2-72B, trained on data in 27 additional languages besides English and Chinese.

Small Is the New Big

SLMs or specialised models are going to be the future, alongside generalised models like GPT-4 or Claude 3.5 Sonnet. “For everyday use, an 8B or maybe 70B LLM will suffice. If you don’t want to test a model to the max, you don’t need a SOTA model. For everyday questions, which now appear in all training data, smaller models are sufficient,” posted a user on X.

“Both specialised and generic models will coexist, one is not a replacement for the other. It’s the wrong dream to believe we only need one API such as OpenAI. We will need both mega models aiming for AGI and specialised micro models that can integrate into today’s workflows,” said Pramod Varma, chief architect of Aadhaar, in a recent interaction with AIM.

In India, Gnani.ai recently launched a series of voice-first SLMs, which are trained on millions of audio

hours of proprietary audio datasets and billions of Indic language conversations, capturing the rich diversity of dialects, accents, and linguistic nuances prevalent across the country.

Abacus.AI chief Bindu Reddy has predicted that in the next five years, smaller models will become more efficient, LLMs will continue to become cheaper to train, and LLM inference will become widespread. “We should expect to see several Sonnet 3.5 class models that are 100x smaller and cheaper in the next one to two years.”

Earlier this year, Microsoft unveiled Phi-3-Mini, a 3.8 billion parameter language model trained on an extensive dataset of 3.3 trillion tokens. Despite its compact size, the Phi-3-Mini boasts performance levels that rival larger models such as Mixtral 8x7B and GPT-3.5.

On similar lines, Karpathy explained that the reason current models are so large is that the training process is still quite wasteful: “We’re asking them to memorise the internet, and, remarkably, they do and can, e.g., recite SHA hashes of common numbers or recall really esoteric facts.” “But imagine if you were going to be tested, closed book, on reciting arbitrary passages of the internet given the first few words. This is the standard (pre)training objective for models today,” he added. He noted that the ability to demonstrate thinking (e.g., reasoning) is closely tied to the amount of knowledge the model has learned from the training data.

He further explained that in the future, as larger models help refine and optimise the training process, smaller models will emerge. “The models have to first get larger before they can get smaller because we need their (automated) help to refactor and mould the training data into ideal, synthetic formats.”

Cost Effective

The constant use of LLMs also has its drawbacks, including high inference costs and energy consumption. SLMs, such as Gemma or LLaMA 8B, are very cost-effective as they require less compute and their inferences consume less energy.

OpenAI used to spend around \$700,000 daily to run ChatGPT. With GPT-4o mini, that cost may decrease to some extent. Moreover, since SLMs can run locally on devices without requiring the cloud, they don’t need extensive AI infrastructure and GPUs.

Compared to Claude 3.5 Sonnet and Google’s Gemini 1.5 Pro, OpenAI’s GPT-4o mini is significantly cheaper. According to Hugging Face, GPT-4o

mini is the least expensive, with a combined input and output cost of just \$0.75 per million tokens.

Ola Krutrim chief Bhavish Aggarwal recently pointed out that a single H100 NVIDIA GPU consumes 30 times more electricity in a year than an Ola S1 scooter. “An H100 GPU requires around 8.7 MWh of energy per year, whereas an S1 requires 0.25 MWh/year. Need a lot of electricity in the future!” he said.

However that is now changing as Karpathy recently demonstrated that today one can train GPT-2 on their own for ~\$672, running on one 8XH100 GPU node for 24 hours. “Incredibly, the costs have come down dramatically over the last five years due to improvements in compute hardware (H100 GPUs), software (CUDA, cuBLAS, cuDNN, FlashAttention) and data quality (e.g. the FineWeb-Edu dataset),” he said.

On the other hand, TWO AI chief Pranav Mistry claimed that the future is in 1-bit LLMs as they have multiple advantages, meaning that they can fit on smaller chips, require less memory, and have faster processing.

SLMs, the future of generative AI?

One question haunts enterprises: the balance between the impressive capabilities of SLMs and their potential drawbacks compared to private instances of large models, and whether small models represent the future of generative AI. “Technology is rapidly evolving how we operate and train. Three months ago, using an SLM isolated in a customer’s private database might have seemed like a drawback compared to a private instance of a large public model,” said Upstage’s Kasey Roh, on the sidelines of the recent AWS Summit in Washington DC.

The South Korean company has built an SLM called Solar, one of the top-ranked models on the HuggingFace Open LLM leaderboard, and a finetune of Llama 2.

Further, Roh said that to address the challenges of SLMs, the company has now entered into a subscription model with its customers, allowing them to continuously train and fine-tune models with the latest data points.

She believes that this modular and standardised approach significantly mitigates the drawbacks of using small private models. “Once we have a more complete set of the tools and programmes available, I think that the drawback of having a small model that’s private can be largely addressable,” concluded Roh. 



Join the World's Leading Data Science Professional Community

Join the World's Leading Data Science Professional Community
ADaSci, the premier global professional body of data science and
machine learning professionals, is committed to providing a
platform where data science professionals can come together,
share their knowledge and ideas and help community members
achieve milestones in the field.



BECOME A MEMBER

www.adasci.org

Why AI is Incredibly Smart and Shockingly Stupid

Yejin Choi, a leading computer scientist, is deeply obsessed with instilling true common sense in AI to prevent catastrophic failures, arguing that while AI's advancements are remarkable, building advanced models isn't for everyone—because "AI today is unbelievably intelligent and then shockingly stupid."

AI startups are not easy to build. The costs of computing alone are astronomical, with a majority of seed funding going towards training models and research. But what if you didn't need to worry about renting GPUs?

Allen Institute of AI's commonsense AI senior research director Yejin Choi posed this possibility at Databricks' Data + AI Summit this year when discussing the possibilities of working with small language models (SLMs). "Right now, the recipe is, 'Let's just make the models super big – the bigger, the better,' but humans – you and I – can't really remember all that context, you know?" she said.

In posing several 'mission impossibles' in building SLMs, one of the constraints she gave was in building a model without a GPU.

Choi's Mission Impossibles for SLMs

The first "mission impossible" scenario that Choi went into was being able to summarise sentences without the use of reinforcement learning from human feedback (RLHF), extreme scale pre-training and supervised datasets at scale.

Taking it a step further, her second "mission impossible" was to do the same, but summarising documents this time instead of sentences, with the additional constraint of not relying on human-supervised critics. Choi reduced the capabilities that her team were working with in the third "mission impossible", stating that they tried to make

“
Please do not
overgeneralise to conclude
that all SLMs are completely
out of the league.
There are numerous
other counterexamples
that demonstrate that
task-specific symbolic
knowledge distillation can
work across many different
tasks and domains.”



Yejin Choi

Wissner-Slivka Chair of Computer
Science, University of Washington.

older statistical n-gram language models relevant to neural language models.

In doing so, she said that they made an equivalent to infinity, it had to compute over trillions of tokens with an almost instantaneous response time, and finally, all of this had to be done without the use of GPUs.

This is done because SLMs do not require the massive amounts of compute that LLMs require, thereby foregoing the need for GPUs. Choi highlighted this by taking the example of Infini.gram, an engine built by researchers from Washington University and the Allen Institute for AI, which can be run with basic compute from CPUs.

In doing so, Choi demonstrated that it was possible to compete with larger models even under previously thought impossible constraints and with limited resources.

Her biggest contention was that AI would be as good as the data it was trained on, hence she speculated that synthesised data would be the way to go in the future. “Please do not overgeneralise to conclude that all SLMs are completely out of the league. There are numerous other counterexamples that demonstrate that task-specific symbolic knowledge distillation can work across many different tasks and domains,” she said, highlighting that SLMs were still plenty effective if certain criteria were fulfilled.

High-Quality SLMs

Choi said that the commonly held belief at the moment is that we aren’t able to efficiently build the human-like ability to abstract information into AI models. “You just abstract away everything I told you instantaneously but you still remember what I said so far. That’s really amazing human intelligence that we don’t yet know how to



“

When you have really high-quality data, or textbook quality data, synthesised, you can actually compete against larger counterparts across many different tasks”

build efficiently through AI models. I believe that it’s possible, we’re just not trying hard enough because we’re blinded by just the magic of a scale,” she said.

This means that, in order to build an efficient SLM, the focus would need to be on perfecting abstraction within the model, rather than increasing the amount of training data that the model relies on.

Synthesised Data

Additionally, Choi focused on synthesised data, rather than what was already available, because, in her words, “We have to synthesise data because if it already exists somewhere on the internet, OpenAI has already crawled it.”

She said that while certain concerns exist about the use of synthetic data, including the quality of the data, as well as potential bias, the actual synthesis of data needs to be done carefully and in a way that is innovative.

Further, she said this was already occurring, highlighting a research paper on annotating synthesised data by Meta AI called the Segment Anything Model (or SAM). Another paper she highlighted was Microsoft’s ‘Textbooks Are All You Need’. “When you have really high-quality data, or textbook quality data, synthesised, you can actually compete against larger counterparts across many different tasks,” she said.

Relying on synthesised data, SLMs can lower the cost of acquiring data and satisfy the need to use novel data while still training the model itself on high-quality data, as demonstrated by the ‘Textbooks Are All You Need’ research paper.

With synthesised data and perfecting abstraction, Choi believes that building SLMs can substitute the need for LLMs, especially thanks to the respite it allows in terms of compute. 

Elevating GenAI Skills Across the Board

Equip your team for any challenge with our comprehensive Generative AI training solution.



We know that talent is
your greatest asset.

Our corporate training program on Generative AI provides a unique opportunity to empower, retain, and advance your talent.



Scan for
more details

15%

Growth in employee
skill Adaptability

18%

Rise in Creative
Solutioning through
AI Integration

22%

Expansion in
Data-Driven Decision
Making Proficiency

By AIM Editorial

Generative AI in Healthcare & Life Sciences in India

Dr Illing completed his undergraduate paediatric training at St John's Medical College and Hospital, Bengaluru, almost two decades ago.

Generative AI holds significant promise for India, given its cultural and language diversity. At the AWS Summit, Washington DC, Rowland Illing, the director and chief medical officer for global healthcare and non-profits, AWS, acknowledged the same. "The appetite for generative AI in India is high. The technology can help with language and culturally sensitive information, addressing the breadth of languages in India. It's exciting to see opportunities in the National Digital Health Mission as well," he told AIM.

Illing's sentiments resonate well since many Indian healthcare companies are actively tapping into the potential of generative AI. Today, every health company is turning into an AI company.

For example, Apollo partnered with Google Cloud to improve its AI-powered Clinical Intelligence Engine (CIE) using MedLM and retrieval-augmented generation (RAG), integrating extensive clinical data for improved decision-making.

While Narayana Health, in partnership with Microsoft, is making diagnostic accuracy and operational efficiency better, Manipal is personalising cancer care through imaging and predictive analytics with AI.

On the startup side, apart from Qure.ai, Khosla Ventures-backed Healthify provides custom diet and fitness plans via its generative AI assistant Ria. The team uses solution models based on diverse statistical models, including AWS, Meta, and OpenAI. "I have a great



I have a great respect for the Indian health-tech ecosystem. I think it's an incredibly rich and exciting space to be in. So, we've got great health-tech companies that are already building on AWS across India," said Illing, emphasising the work of Bengaluru-based startup Qure.ai in medical imaging AI.

Dr Rowland Illing

Director and Chief Medical
Officer for Global Healthcare
and Non-Profits, AWS



respect for the Indian health-tech ecosystem. I think it's an incredibly rich and exciting space to be in. So, we've got great health-tech companies that are already building on AWS across India," said Illing, emphasising the work of Bengaluru-based startup Qure.ai in medical imaging AI.

AWS is closely working with many widespread Indian companies, such as the Center for Cellular and Molecular Biology in India, the Poshan Tracker, and the CoWIN vaccination platform. During the pandemic, CoWIN was built natively on AWS, supporting 2.2 billion vaccination doses across the country.

Beyond the functioning of AWS in India, Illing also shares another special connection with India. The Oxford University alumni completed his undergraduate paediatric training at St John's Medical College and Hospital, Bengaluru, almost two decades ago.

Amazon Bedrock for Health

Meanwhile, Genomics England, a UK government-owned entity, is leveraging Claude on Amazon Bedrock, AWS' fully managed generative AI service, to assist researchers in linking genetic variants to medical conditions. Illing told AIM that it resulted in the discovery of 20 new gene associations with learning disabilities, providing actionable insights for future research and treatment.

This research, based on peer-reviewed articles, aims to improve genetic tests and health, especially for

intellectual disability, by quickly processing vast data to identify potential gene associations.

Genomics England used Anthropic's LLM Claude, available through Amazon Bedrock, to make this happen. Interestingly, Amazon concluded a \$4 billion investment in the AI startup in March of this year. "We have a great relationship with Anthropic. We're invested in them, and we get early releases of their models, like Claude 3.5 through Bedrock," he added. "Bedrock is a great vehicle because it allows users to access foundation models through APIs and provides functionalities like guardrails to ensure LLM answers are bound by constraints. Users can enrich foundation models with their data without sharing it with the model."

However, this is just one example, as numerous companies leverage Bedrock's diverse models from various AI providers like Anthropic, OpenAI, and Meta, to meet their unique needs, supporting AWS' mission to democratise foundational models.

The model evaluator function allows users to compare the models based on speed, cost, performance, or accuracy. "We believe there won't be one foundation model to rule them all but hundreds, if not thousands, giving users a choice in a secure way," commented Illing. "The current trend we are observing is the explosion of analytics and generative AI. This is an end-to-end data story, focusing on extracting meaningful insights from



The current trend we are observing is the explosion of analytics and generative AI. This is an end-to-end data story, focusing on extracting meaningful insights from vast amounts of data, often multimodal. Generative AI plays a crucial role in understanding and analysing this data.”

vast amounts of data, often multimodal. Generative AI plays a crucial role in understanding and analysing this data.”

At the AWS Summit Bengaluru in May, the company announced that Amazon Bedrock is now generally available in the Mumbai (Asia Pacific) region to allow large organisations across India to build and scale generative AI applications with enterprise-grade security and privacy.

Since 2017, AWS, the cloud subsidiary of Amazon, has trained over 5.5 million people in India on AI and cloud skills and over 8.3 million across the APAC and Japan regions.

With generative AI moving from proof of concept to deployment this year, the company plans to train about two million individuals in AI globally by 2025 through its “AI-Ready” initiative.

Shifting to Cloud

Apart from generative AI, there is a global shift happening towards cloud adoption in healthcare, especially after the COVID-19 pandemic, which needed remote work. Despite a decrease in global IT spending in 2020, cloud spending grew by over 6%, reaching \$258 billion, with experts predicting the market will double in the coming years. “Previously, the cloud was used for disaster recovery and backup, running things on-premises. However, with security being a top priority and concerns about outages, bad actors, and ransomware attacks, we are seeing a shift in organisations moving their production clinical environments to the cloud,” explained Illing.

This involves using clinical workloads with patient-identifiable data in a cloud-read-only mode initially, before running things on-premises. Over time, organisations have found that cloud-based systems often outperform their on-premises counterparts, leading to a complete transition to cloud-based production environments.

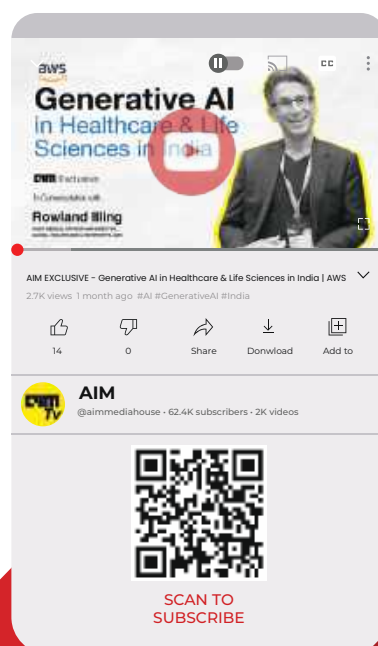
As per Nutanix’s report, the future of healthcare relies on moving away from legacy systems. Currently, 27% of

healthcare companies use only traditional, non-cloud data centres and business systems, more than any other industry.

Several notable examples illustrate this trend. The world-renowned academic medical centre, Tufts Medicine and Danville-based Geisinger Health in the US, now run its electronic healthcare records (Epic) natively on AWS.

Similarly, New South Wales in Australia is migrating its entire eHealth system, including the single digital patient record for the state, onto AWS. Apart from new startups, legacy health giants like Pfizer, Amgen, and Merck, are also building drug discovery platforms on AWS.

More recently, AWS chartered into protein folding. It has partnered with the NY-based research team EvolutionaryScale to bring the latter’s ESM3 multimodal large language models available on AWS, aiding applications from drug discovery to carbon capture. 



By AIM Research

Top Data Science Service Providers – PeMa Quadrant 2024

AIM Research's annual initiative to recognise leading Data Science Service Providers sets an industry standard, helping businesses quickly assess the potential of vendors.



The data science industry has become a critical driver of innovation and business growth across several sectors. As organisations increasingly rely on data-driven insights to guide strategic decisions, the demand for specialised data science service providers has surged. To assess and identify leading service providers in the data science industry, AIM Research conducts an annual survey in which vendors participate voluntarily to benchmark their capabilities against their peers. **A total of thirty-six service providers participated in this year's assessment.** All the vendors featured in this report are analysed through a survey, which gathered quantitative and qualitative data on various aspects of their operations, such as financial health, growth, customer confidence, client reach, work delivery, tech advancement, employee maturity, and support infrastructure.

Through this analysis, we aim to show how the industry has advanced over the past year and highlight the top data science service providers, with firms expanding their teams, enhancing their expertise, and forging strategic partnerships to stay competitive in a rapidly evolving market.

Key Developments in the Space of Data Science Service Industry

Expansion of GenAI Service Portfolios: Data science service providers are expanding their offerings beyond traditional data analytics to include advanced services such as AI and ML, deep learning, computer vision, and GenAI. Many service providers have introduced GenAI accelerators to quickly deliver advanced AI-driven analytics solutions and stay competitive in the evolving market.

Emphasis on Industry-Specific Solutions: There is a trend towards specialization for developing industry-specific solutions that cater to the unique needs of sectors such as BFSI and healthcare and life sciences. Providers focusing on the financial services sector are developing advanced risk assessment models, while those in healthcare are creating predictive analytics tools for patient outcomes.

Demand for Cloud Native Solutions: There is a strong shift towards cloud-based and hybrid solutions, driven by the need for scalability, flexibility, and remote accessibility. Many providers are integrating their services with leading cloud platforms such as AWS, Google Cloud, and Microsoft Azure, offering clients the ability to scale their data science initiatives efficiently.

Data Governance & Synthetic Data: Service

providers are implementing policies and practices to address issues such as bias in AI models, data privacy, and compliance with regulatory standards like GDPR. With the rising concerns around data privacy, the creation of synthetic data, which can mimic the statistical properties of real data without containing any actual information, will become prominent.

Focus on Upskilling & Reskilling: Vendors are focusing on upskilling and reskilling their teams in emerging areas like AI/ML, GenAI, cloud computing, and big data analytics. They are also investing in training their teams on industry-specific knowledge, especially in sectors like healthcare, finance, and retail, to enhance domain expertise.

Partnerships and Alliances: Strategic partnerships and alliances are becoming a key component of growth strategies for data science service providers. These partnerships often involve collaboration in the areas of AI and cloud optimisation with hyperscalers, academic institutions, and industry-specific leading organizations to enhance capabilities and extend market reach.

PeMa Quadrant

We measure the success of service providers in two key areas—Penetration and Maturity. By plotting their relative positions on the PeMa Quadrant, it reveals each company's unique approach to delivering world-class services.

Assessment

The PeMa Quadrant assessment for 'Data Science Service Providers 2024' is conducted based on the participants' responses to the below parameters.

Penetration (Pe)

Financial Health: Analytics workforce, Revenue, Projects, Clients.

Growth: YoY growth in data science revenue and data science work force.

Customer Confidence: Fortune 500 clients, billing structure.

Company Outreach: Office locations, client locations, industries/markets

Maturity (Ma)

Work Delivery: Advanced analytics, data compliance standards, competitive advantage.

Tech Advancement: Scalability, changes in the field and company's approach.

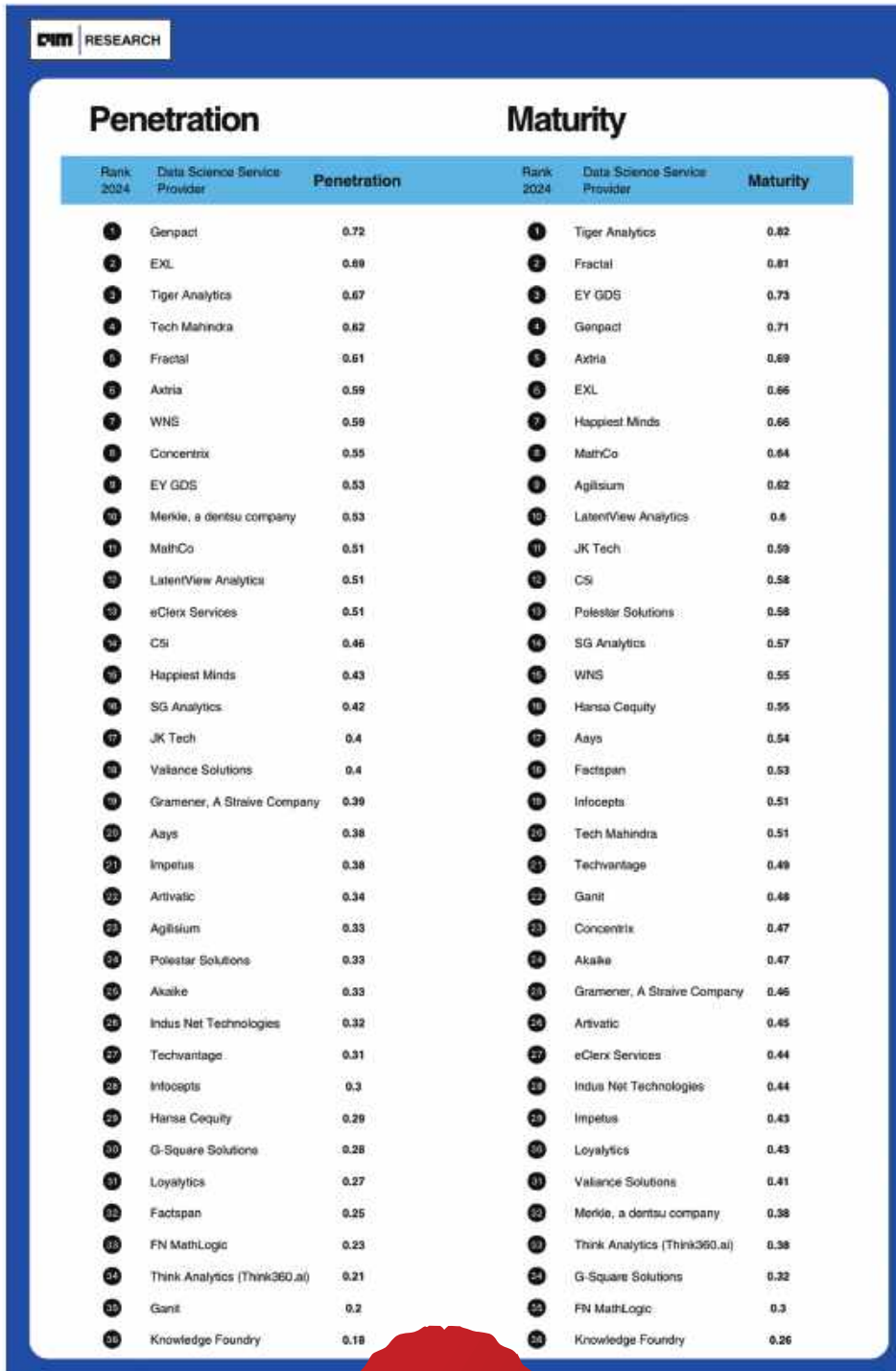
Employee Maturity: Attrition rate, experience, average tenure, performance metrics.

Support Infrastructure: R&D, workforce training, alliances/partnerships.



Copyright © 2024 AIM Media House LLC. All rights reserved.

How has the previous Data Science PeMa Quadrant helped various stakeholders in the ecosystem?



Scan to Download
Full Report

Industry Standard: AIM Research is recognised as a leading analyst firm in the data and AI fields. The annual PeMa initiatives are considered an industry standard and help businesses quickly assess the potential of vendors.

M&A: Several organisations have reached out to inquire about the vendors, particularly those positioned in the challengers quadrant. We conduct in-depth studies to analyse vendors (at their discretion) and connect them with the right players for further collaborations.

Benchmark Capabilities: Participants have the opportunity to compare themselves against peers in the market and identify future growth opportunities. The survey's PeMa questions also serve as a framework for service providers to uncover insights about their businesses.

Vendor Recognition: The PeMa Quadrant's findings can be used in various forms of communication to enhance market reach. It's not just the leaders, growth vendors, and seasoned vendors who gain recognition; challengers, with their smaller size and scope of services, possess an advantage that rivals can't match—superior training and delivery capabilities. 

Why This Semiconductor Chip Company Chose India Over China

“India will be an AI product focussed leader globally and a lot of innovation in AI will come from the country.”

Growing geopolitical tensions, coupled with supply chain constraints have pushed many countries, including India, to consider building their own semiconductor capabilities. Over time, the US has attempted to hinder China's semiconductor capabilities and the latter has managed to find ways to tackle US sanctions. “I really think chips are the new oil. While oil defined the 20th century, chips will define the 21st century and every country will want to own elements of manufacturing,” Krishna Rangasayee, the CEO and founder of SiMa.ai told AIM in an exclusive interview.

Founded in 2018, the startup positions itself as a software company which also builds its own silicon. It focuses on developing AI solutions for the embedded edge computing market.

Despite being headquartered in San Jose, California, SiMa.ai has a significant presence in India and around half of the innovation the startup does happens in India.

Interestingly, the company has chosen India as a strategic market, over China, which is the largest semiconductor market in the world.

Engaging with India, Not China

Rangasayee's recent visit marked his third trip to India in a year. During these visits, Rangasayee, who has accumulated over 18 years at Xilinx, dedicates much of his time engaging with a diverse array of customers, including automobile manufacturers, drone companies, robotics firms, and public sector entities.

The company is betting on India's renewed focus on manufacturing and semiconductor fabrications ambitions.

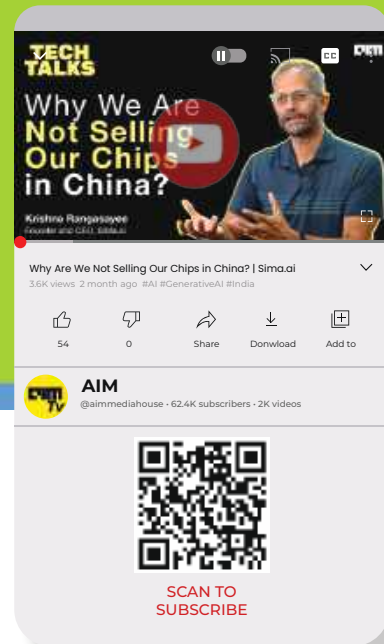
“I think India is going to be an AI product focussed leader globally and a lot of innovation in AI will come from India,” he said.

However, for strategic reasons, the company has chosen not to sell its chips in China. “The world of AI and chips and geopolitics is going to be complicated. We are a startup and we are very focussed on the applications that are coming together in the US and Europe. I have no doubt that it will be happening in India as well,” said Rangasayee.

It's a bold move for a startup, especially considering China's status as the world's largest semiconductor market. In 2022, global semiconductor sales totaled \$574 billion, with China capturing a significant \$180 billion, equating to a 31.4% market share.

Rangasayee's decision to not engage with China might also be influenced by the US restrictions that prevent US companies from directly selling chips to China. This was done to limit China's access to high-end graphics processing units (GPUs) developed by NVIDIA.

On the contrary, China has committed to fulfilling nearly 70% of its domestic silicon requirements internally by 2025. Rangasayee also said his decision to not engage with China has to do with AI regulation. “AI is a complicated area. At some point, there will be an interesting set of regulations that come together and as a startup, we need to be careful. We are not a big company and we cannot afford to engage and then revisit our engagement,” he said. “We have so many opportunities in the US, Europe, India, Korea and Japan and that's a very large market footprint for us,” he added.



Preparing for a Multimodal World with Second-Gen Chips

The startup also recently announced its second-generation chips, which are fabricated by the Taiwanese semiconductor giant TSMC, are under preview and will be publicly launched in the first quarter of 2025.

Built on 6 nm nodes, these chips are designed to run generative AI models on the edge along with computer vision models. The initial generation of these chips is tailored for computer vision applications at the edge only. “The gen-one chips were on a 16 nm technology and no doubt we will have improvements in terms of power primarily but also on performance. Our partners continue to be TSMC, Arm and Synopsys and we are really appreciative of the help they provide in building this platform,” he said.

The startup aims to build one platform catering to original equipment manufacturers (OEMs) running convolutional neural networks (CNNs), vision transformers, as well as large multimodal models (LMMs).

Rangasayee believes multimodal is the future and this is what the startup has been preparing for. “Multimodal is going to be everywhere, from every device to appliances, be it a robot, or an AI PC. You will be able to converse, watch videos, parse inputs, just like you talk to a human being.”

The startup has also outlined plans for a third generation of chips. Rangasayee mentioned the company’s potential to sell its chips to phone and PC makers, but emphasised a cautious approach. However, the startup has no intentions to develop chips for data centres.

AGI Might Not Happen in Our Lifetime

While Rangasayee envisions an AI-dominated future, we did ask his thoughts on the current discourse on artificial general intelligence (AGI).

While some notable figure has predicted AI to supersede human-level intelligence within this decade, Rangasayee has a different outlook.

He believes the current AI systems are nowhere near sentience. “I don’t see that happening with the current form of AI, at least in the next 50–100 years. I don’t see a rationale for how that’s connected. “A lot has been written about it but what is happening is we are moving into an element of multimodality and mimicking things that really help us but the underlying architectures are far (x3) away from being sentient.”

Moreover, many in the industry hold the media complicit in hyping the capabilities of the current AI systems. Rangasayee agrees, “Many smart minds have already spoken about it, but unfortunately the media hypes the capabilities of AI.

Nonetheless, he also adds that, “You can never say never. I am an engineer at heart and I live this everyday. I don’t see a roadmap for AI systems having better human capacity and capability.”

| By Amit Raja Naik

THE FUTURE IS AI AGENTS. NO DOUBT | 44

THE FUTURE IS AI

AGENTS. NO DOUBT.

Everybody is talking about them. In fact, no conversation in tech circles goes without the mention of 'AI agents' or 'agent AI' or 'compound AI systems' these days.



“I am actually convinced that the adoption of agent AI will be faster,” said CP Gurnani, the former CEO of Tech Mahindra and the co-founder of AIONOS, at our MachineCon GCC event, held recently in Namma Bengaluru. AIONOS is an AI startup, which Gurnani started with Rahul Bhatia, the group MD of InterGlobe Aviation (IndiGo).

Gurnani envisions a future where each of us will have a personalised AI avatar. The avatar will be deeply familiar with our daily tasks and activities, enabling it to analyse the business and rule engines and provide invaluable support to enhance productivity and efficiency. Check out the full interview, exclusively on AIM.

In the past few months alone, there has been a body of research surrounding AI agents. Andrew Ng, too, is obsessed with it and has been busy releasing a series of courses on how to build one.

Developers’ favourite Andrej Karpathy has also been advocating agent OS for some time now, speaking about how it will transform industries through its ability to automate and optimise complex workflows rapidly and efficiently. “With agent OS, you’re able to create multi-agent workflows in a matter of seconds from natural language,” he added.

So, what are AI agents?

Simply put, AI agents are systems that perform complex tasks and make decisions, thereby enhancing productivity, efficiency, and innovation across industries.

At present, AI agents are being used in various sectors, including coding, healthcare, finance, and legal, and we are already starting to see some early signs of their applications in the real world.

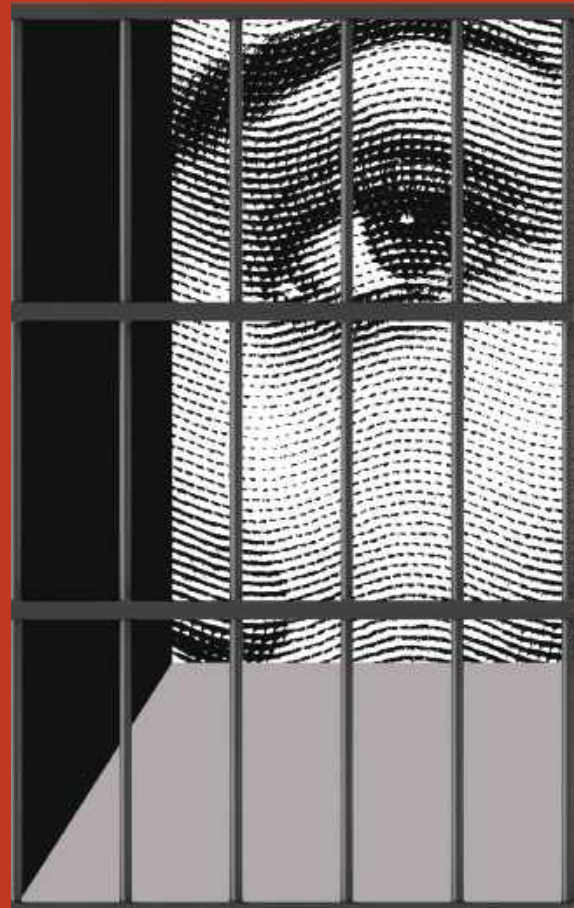
However, it is easier said than done.

“Hallucinations block a lot of enterprise use cases,” said Dario Amodei, the CEO of Anthropic, in a fireside chat with entrepreneur, investor and startup helper Elad Gil, at GoogleCloudNext, adding that Anthropic is working really hard to reduce the hallucinations present in models. “What I expect to see in the next year or two are agents,” he added, saying that, so far, we’ve seen lots of instantiations of agents and not seen anything that is really, really strong. “I think getting something that’s really strong will need additional scale and techniques on top of that,” said Amodei.

He said the models need to get more reliable because the individual steps must have very low error rates. And a part of that will come from scale. “We

“What I expect to see in the next year or two are agents,” he added, saying that, so far, we’ve seen lots of instantiations of agents and not seen anything that is really, really strong. “I think getting something that’s really strong will need additional scale and techniques on top of that.”

—Amodei.



AI AGENTS, EVERYWHERE!

AI agents are set to transform industries in the coming months by seamlessly integrating advanced automation and decision-making capabilities, driving unprecedented levels of innovation, efficiency, and productivity, and unlocking new possibilities across various sectors.

CODING	Cognition Labs (Devin) Stition AI (Devika) Replit Code Repair OpenDevin
HEALTHCARE	The Children's Hospital of Philadelphia (D.A.V.I.D) Institute for Future Health (CHA)
FINANCE	INATIGO (Finley AI) AI4Finance Foundation (FineRobot)
LEGAL	Harvey ContractKen

BUILD YOUR OWN AI AGENTS (BYOAIA)

Open

AutoGen

fabric

crewAI

SuperAGI

Langroid

AI

ADEPT

cohere

closed

AIM Media House © Copyright 2024. All rights reserved.

need another generation or two of scale before the agents really work,” he added.

Amodei said that as models start to become agents, as they go out in the physical world, and as they take actions on computers, phones, and interfaces autonomously and without humans overseeing them, there will be very substantial risks in this area —“like what guardrails; how do I know this thing didn’t introduce a security vulnerability intentionally or by accident?”

Autonomous AI agents will have to wait. Microsoft AI chief Mustafa Suleyman said that it won’t be until GPT-6 in 2 years’ time that AI models will be able to follow instructions and take consistent action. “I think it’s going to require not just one but two orders of magnitude more computation to train the models.”

As OpenAI CTO Mira Murati said in an interview, the next generation of GPT will be ‘PhD-level’ compared to GPT-3 (toddler) and GPT-4 (high school). She also said the next model will be released in a year and a half.

Early Signs of Autonomous AI Agents

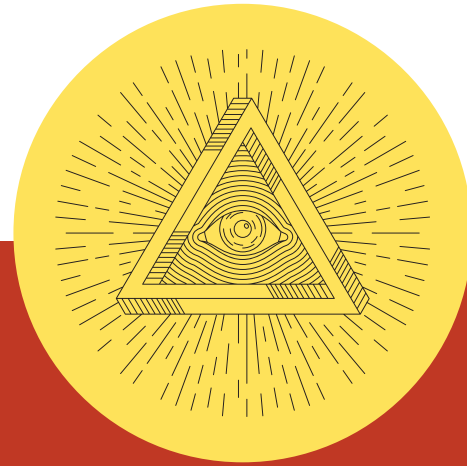
“A lot of them are basically using humans in the loop to help steer something. So an agent is almost like a junior employee or intern who’s coming back and asking you for help,” quipped Amodei. This includes Cognition Labs (Devin) and alike for coding, and Magic.AI by YC-backed Fieldprox, which optimises field service management by intelligently automating task assignments based on performance metrics and geographical proximity.

Recently, Children’s Hospital of Philadelphia (CHOP), in collaboration with AWS, announced the launch of D.A.V.I.D., a generative AI-powered navigator designed to assist clinicians and researchers in real time with questions about clinical trials, patient data, treatments, and outcomes.

A few months ago, the Institute for Future Health unveiled Conversational Health Agents (CHA), a personalised LLM-powered framework designed to assist healthcare professionals with real-time conversational AI capabilities.

Similarly, in finance, the AI4Finance Foundation released FinRobot, an open-source AI agent platform designed for financial applications, which uses LLMs to automate complex financial tasks and enhance decision-making.

In legal, besides Harvey, which is backed by OpenAI, ContractKen, started by former EXL exec Amit Sharma in 2021, is building the world’s best AI copilot for contracts, assisting legal professionals in corporate legal teams and law firms to draft, review, and negotiate contracts efficiently. 



“

A lot of them are basically using humans in the loop to help steer something. So an agent is almost like a junior employee or intern who’s coming back and asking you for help.”



The 'Linux Moment' in AI Has Finally Arrived

Developers can run inference on Llama 3.1 405B on their own infrastructure at roughly 50% of the cost of using closed models like GPT-4o.

A few days ago, AIM asked a pertinent question: “When will the ‘Linux moment’ in AI arrive?” It looks like it has arrived sooner than expected. recently Meta released Llama 3.1, with 405 billion parameters, its largest and best model yet.

The state-of-the-art model outperformed OpenAI’s GPT-4o on several benchmarks, making it the best open-source model available.

Llama 3.1 is competitive with leading foundation models like GPT-4, GPT-4o, and Claude 3.5 Sonnet across various benchmarks. The smaller models in the Llama 3.1 series also perform on par with both closed and open models of similar parameter sizes.

Meta chief Mark Zuckerberg is confident that Llama 3.1 will have a similar impact on the AI ecosystem as Linux had on the operating system world.

“Today, Linux is the industry standard foundation for both cloud computing and the operating systems that run most mobile devices – and we all benefit from superior products because of it,” he said, adding, “I believe that AI will develop in a similar way. Today, several tech companies are developing leading closed models. But open source is quickly closing the gap.” “I think open source AI is going to become the industry standard just

like Linux did. It gives you control to customise and run your own models. You don’t have to send your data to another, and it’s more affordable,” he added.

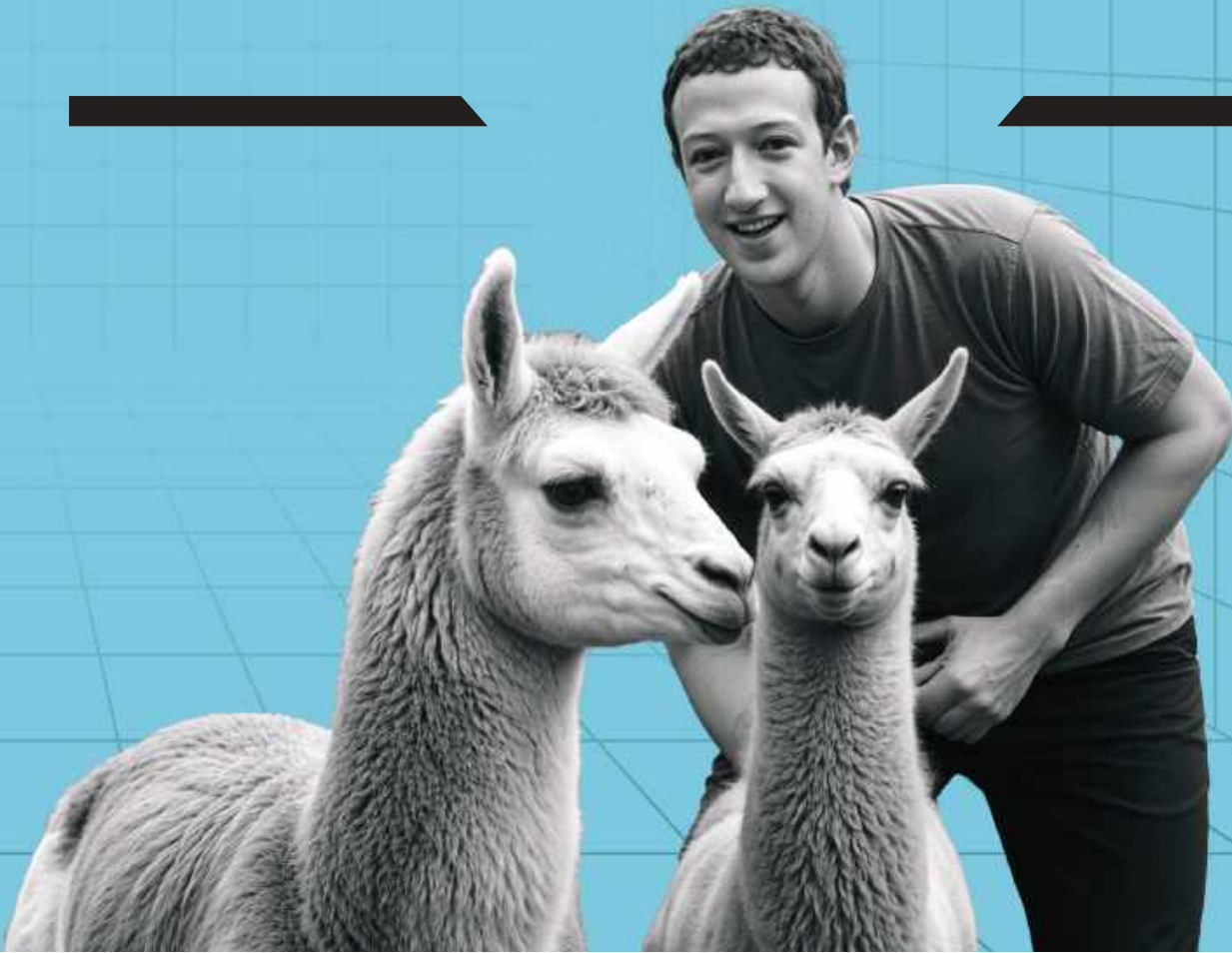
Echoing similar sentiments, OpenAI co-founder Andrej Karpathy said, “I’d like to say that it is still very early days, that we are back in the ~1980s of computing all over again, that LLMs are a next major computing paradigm, and Meta is clearly positioning itself to be the open ecosystem leader of it.”

He elaborated on the applications of AI models, stating that people will prompt and retrieve information from the models, fine-tune them, distil them into smaller expert models for specific tasks and applications, and study, benchmark, and optimise these models.

Recently, Karpathy also discussed how AI kernels could potentially replace current operating systems.

Karpathy envisions an operating system where the LLM acts as the kernel, managing and coordinating system resources and user interactions, and AI agents function as applications, providing various services and functionalities.

In this system, natural language serves as the primary programming and interaction interface, allowing users to communicate with the system in plain English or other languages.



“A lot of what we’re focused on is giving every creator and every small business the ability to create AI agents for themselves, making it so that every person on our platforms can create their own AI agents to interact with.”

AI Agents, Everywhere

Zuckerberg anticipates that Meta AI will become the leading chatbot by the end of this year, given that ChatGPT currently boasts over 100 million users. Interestingly, Meta has not yet disclosed any usage statistics for its own assistant.

Meta anticipates that billions of AI agents will be developed worldwide using open-source models. “Our vision is that there should be a lot of different AIs and AI services out there, not just one singular AI, and that really informs the open-source approach,” said Zuckerberg.

“A lot of what we’re focused on is giving every creator and every small business the ability to create AI agents for themselves, making it so that every person on our platforms can create their own AI agents to interact with,” he added.

Vinod Khosla has been dreaming about it too. He envisions a future in which internet access will mostly be through agents acting for consumers, carrying out tasks, and fending off marketers and bots. “Tens of billions of agents on the internet will be normal,” he wrote. “Eventually, all our interactions with the digital world will be mediated by AI assistants. This means that AI assistants will constitute a repository of all human knowledge and culture; they will constitute a shared infrastructure like the internet is today,” said LeCun.

He urged platforms to be open-source and said that we cannot have a small number of AI assistants controlling the entire digital diet of every citizen across the world, taking a dig at OpenAI and a few other companies without naming them. “This will be extremely dangerous for diversity of thought, for democracy, and for just about everything,” he added.

Surprisingly, for the first time in history, Meta also updated its licence to allow developers to use the outputs from Llama models — including 405B — to improve other models.

“We’re excited about how this will enable new advancements in the field through synthetic data generation and model distillation workflows, capabilities that have never been achieved at this scale in open source,” said the company.

Last week, a faulty sensor configuration update by CrowdStrike caused a significant Microsoft Windows outage, impacting global transport, finance, and medical sectors. Looking ahead, if AI agents are managed by a single entity and cloud infrastructure, a similar failure could re-occur.

“I think the main thing that people are going to do [with Llama 3.1 405], especially because it’s open source, is use it as a teacher to train smaller models that they use in different applications,” said Zuckerberg. “If you just think about all the startups out there, or all the enterprises or even governments that are trying to do different things, they probably all need to, at some level, build custom models for what they’re doing,” he added.

The idea was suggested earlier by Karpathy, who explained that in the future, as larger models help refine and optimise the training process, smaller models will emerge. “The models have to first get larger before they can get smaller because we need their (automated) help to refactor and mould the training data into ideal, synthetic formats.”

Last week, we saw the release of several small models that can be run locally without relying on the cloud. Small language models, or SLMs, are expected to become the future alongside generalised models like GPT-4 or Claude 3.5 Sonnet.

There is no doubt that OpenAI felt pressure as it made fine-tuning available for its latest model, GPT-4o mini. “Customise GPT-4o mini for your application with fine-tuning. Available today to tier 4 and 5 users, we plan to gradually expand access to all tiers. The first 2M training tokens a day are free, through Sept 23,” the company posted on X.

Meanwhile, Zuckerberg has claimed that developers can run inference on Llama 3.1 405B on their own infrastructure at roughly 50% of the cost of using closed models like GPT-4o, for both user-facing and offline inference tasks.

Jealous Much?

One thing is clear, OpenAI can’t stay silent when its competitors are on the verge of releasing a new model. Throughout the day, the company has been emphasising safety and preparedness. “We won’t release a new model if it crosses a ‘medium’ risk threshold until we implement sufficient safety interventions,” the company said. OpenAI appears to be subtly suggesting that open-source models might present risks to society.

Mira Murati, the chief technology officer of OpenAI, during a recent interview at the AI Everywhere event at Dartmouth College, said that OpenAI gives the government early access to new AI models, and they have been in favour of more regulation. “We’ve been advocating for more regulation on frontier models, which will have these amazing capabilities and also have a downside because of misuse. We’ve been very open with policymakers and working with regulators on that,” she said.

On the other hand, many are concerned about potential regulations stifling innovation. “I hope foolish regulations like California’s proposed SB1047 don’t stop such innovations,” said DeepLearning.ai founder Andrew Ng on Llama 3.1 release.

Meanwhile, Meta has decided not to release its multimodal Llama AI model in the European Union due to regulatory concerns.

The company cited the lack of clarity and predictability in the EU’s regulatory framework for AI, which includes existing regulations like the General Data Protection Regulation and forthcoming ones such as the AI Act.

The company’s focus on developing its Language Processing Unit (LPU) is pivotal in addressing the computational needs of modern AI models. The advantages of Groq’s hardware lie in its ability to deliver high-speed processing with exceptional energy efficiency and predictable performance.

The LPU architecture is designed specifically to excel in AI workloads, including those intensively demanding models like Llama 3.1. This integration results in ultra-low latency inference when deploying Llama models in cloud environments. ■



NIMHANS
CONVENTION CENTRE
BENGALURU

FEBRUARY | 5 TO 7, 2025

MACHINE LEARNING DEVELOPERS SUMMIT

2025

7TH
Edition

2000+
Attendees

100+
Speakers

600+
Organizations

THE BIGGEST AI CONFERENCE FOR
DEVELOPERS IN INDIA



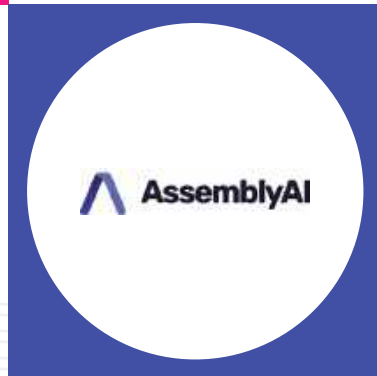
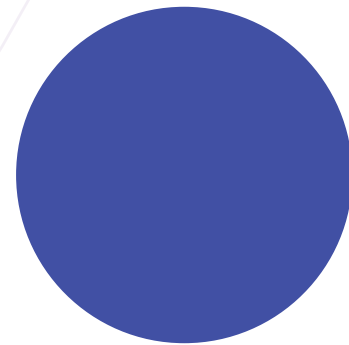
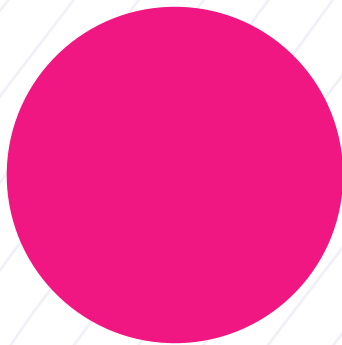
www.mlds.analyticsindiamag.com

FASTEST GROWING GENAI STARTUPS

to Watch Out for in the US 2024

The generative AI market is experiencing explosive growth, with the global market size expected to reach \$66.62 billion by the end of 2024, up from \$44.89 billion in 2023 – a staggering 48.4% year-over-year increase. This rapid expansion is driven by widespread adoption across industries, with 92% of Fortune 500 firms already implementing generative AI solutions.

As the market continues to evolve at a breakneck pace, several US-based startups are emerging as frontrunners in the generative AI race. Here are some of the fastest-growing generative AI startups to watch in 2024:



10

Suno

Founded in 2022 by Mikey Shulman as CEO, Suno revolutionizes music creation with AI-generated songs from text prompts. Despite legal challenges, the company has secured significant investment and partnerships with major tech companies. Shulman, with a PhD in physics from Harvard and experience at MIT Sloan School of Management, leads Suno's mission to democratize music creation while addressing copyright concerns.



09

EvolutionaryScale

Founded in 2023 by Alex Rives as CEO, EvolutionaryScale develops AI for biology, including the ESM3 protein language model. With \$142 million in seed funding, Rives' experience from leading Meta's AI protein team drives the company's mission to revolutionize drug discovery and environmental science.



EvolutionaryScale



08

Atomic AI

Founded in 2021 by Raphael Townshend, Atomic AI is at the forefront of RNA drug discovery. The company's platform uses deep learning to discover and design RNA-targeted small molecules and RNA-based medicines. With \$35 million in Series A funding, Townshend's background in AI and structural biology drives Atomic AI's mission to unlock new treatments for previously undruggable diseases.



07

Prosimo

Founded in 2019 by Ramesh Prabakaran as CEO, Prosimo revolutionizes multi-cloud infrastructure with its Application Experience Infrastructure (AXI) platform. Under Prabakaran's leadership, the company leverages public cloud infrastructure to simplify and secure multi-cloud application delivery. With recognition such as the 2021 Networking Trailblazers Award, Prosimo is shaping the future of cloud networking and security.





06

Observe.AI

Founded in 2017 by Swapnil Jain, Observe.AI enhances contact center performance with its AI-driven conversation intelligence platform. The company analyzes 100% of customer calls, providing real-time insights and coaching for agents. With over 150 customers, including Root Insurance and Tripadvisor, and \$213 million in funding, Observe.AI is a key player in transforming customer service operations. Jain, with product experience from Twitter and Google, leads the company's growth.



05

Harvey

Founded in 2022 by Gabriel (Gabe) Pereyra and Winston Weinberg, Harvey focuses on AI solutions for the legal sector. Leveraging OpenAI's GPT technology, the company provides AI-powered tools for law firms and legal departments. Under CEO Pereyra's leadership, Harvey has secured \$80 million in Series B funding, valuing the company at \$715 million. Pereyra, a former DeepMind research scientist and Meta AI engineer, aims to revolutionize legal workflows, with major firms like Allen & Overy and PwC adopting Harvey's solutions.



Harvey.



04

AssemblyAI

Founded in 2017 by Dylan Fox, AssemblyAI is a rapidly growing AI company specializing in speech recognition and natural language processing technologies. As CEO, Fox has led the development of cutting-edge APIs that offer advanced speech-to-text transcription, speaker diarization, and content moderation capabilities. Under his leadership, AssemblyAI has secured significant funding, including a \$30 million Series B round in 2022, and has attracted major clients across various industries. Fox's vision for AssemblyAI focuses on making powerful AI tools accessible to developers and businesses, enabling them to easily integrate state-of-the-art speech AI into their applications and workflows. The company's innovative approach and high-quality APIs have positioned it as a leader in the speech AI market.



03

Hugging Face

Founded in 2016 by Clément Delangue, Hugging Face is a pioneering AI company known for its open-source machine learning platform. As CEO, Delangue has led the development of the widely-used Transformers library, which simplifies the deployment of state-of-the-art NLP models. Under his leadership, Hugging Face has secured significant funding, including a \$235 million Series D round in 2023, valuing the company at \$4.5 billion. Delangue's vision focuses on democratizing AI and fostering a collaborative community, making Hugging Face a central hub for AI researchers and developers worldwide.



Hugging Face



02

Perplexity

Co-founded in 2022 by CEO Aravind Srinivas, Perplexity is a rapidly growing AI-powered search engine and question-answering platform. The company has gained traction with over 10 million monthly active users and \$165 million in funding, valued at over \$1 billion. Srinivas, holding a PhD in Computer Science from UC Berkeley and a former AI researcher at OpenAI, leads the mission to revolutionize online information retrieval. Perplexity's AI provides direct, sourced answers to queries, setting it apart from traditional search engines.



01

Anthropic

Founded in 2021 by former OpenAI employees, including CEO Dario Amodei, Anthropic is a major player in generative AI. The San Francisco startup focuses on developing safe and reliable AI systems, with its Claude family of language models competing against industry giants. Anthropic has attracted significant investments, including up to \$4 billion from Amazon and \$2 billion from Google. Under Amodei's leadership, the company emphasizes AI safety and transparency, using innovative approaches like "AI Safety via Debate" and "Factored Cognition." Amodei, with a PhD in physics from Princeton, brings valuable expertise to guide Anthropic's mission of aligning powerful AI systems with human values.



ANTHROPIC



RESEARCH

Driving Actionable AI & Data Science Insights, Research, and Strategies.

Pioneering AI &
Data Science
Research Since
2012: With over a
decade of
expertise, AIM
Research is your
trusted guide in
the digital frontier.



www.aimresearch.co

BECOME A MEMBER

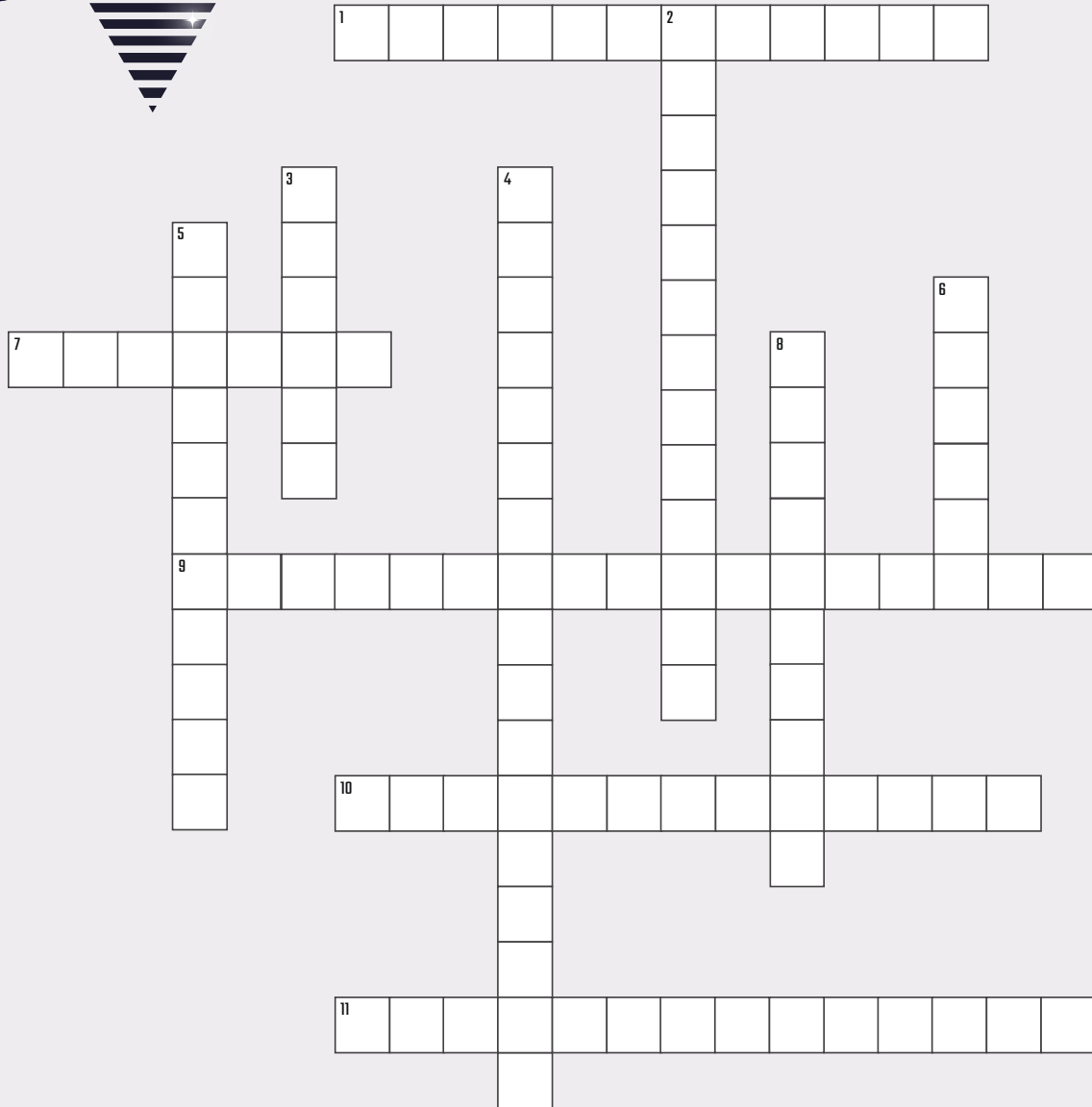
BRAINWAVE ARCADE

AI GAME ZONE

Dive into mind-bending puzzles and challenges that spark
your inner tech genius!



Complete the crossword puzzle below



1. Core of AI research that deals with learning from data
2. Data analysis that involves finding patterns
3. Type of neural network commonly used for image recognition
4. Learning technique uses rewards to reinforce behaviour
5. Type of learning that occurs without labelled data
6. Type of learning that requires labelled data for training
7. Kind of network inspired by the human brain's structure
8. Field of study focusing on analysing complex data sets
9. Type of learning that adjusts itself based on feedback
10. System capable of making decisions without human intervention
11. Programming language that is widely used in AI research and development

1. Machine learning 2. Pattern 3. Convolutional 4. Reinforcement 5. Unsupervised 6. Supervised 7. Neural 8. Data science 9. Adaptive learning 10. Autonomous 11. Python

Answers:

AI Word Quest



FIND THE WORDS HIDDEN IN THE GRID

B	R	O	B	O	T	I	C	S	M	A
I	D	A	T	A	Q	W	E	R	T	Y
N	I	B	M	A	C	H	I	N	E	U
A	L	G	O	R	I	T	H	M	P	I
R	I	C	X	V	A	I	L	E	D	T
Y	F	G	H	J	I	A	L	I	T	L
O	K	O	D	E	C	D	N	O	G	I
T	P	A	S	D	F	I	G	H	J	O
E	K	I	N	T	E	L	L	I	G	E
D	A	T	A	S	D	F	G	H	J	N
D	A	Y	U	I	O	P	L	E	A	R
R	O	B	O	T	L	O	C	I	T	F
A	R	T	I	F	I	C	I	A	L	S

Answers: Robotics, Data, Algorithm, Machine, Artificial, Intelligence, Network, Robot, Code, Binary

Cipher Shuffle

Unscramble the jumbled AI words to reveal their true forms

UCRLTES

--	--	--	--	--	--	--

IAGNTRIN

--	--	--	--	--	--	--	--

FNEICRNEE

--	--	--	--	--	--	--	--

OSGERESRIN

--	--	--	--	--	--	--	--	--

SLCICFATOINSA

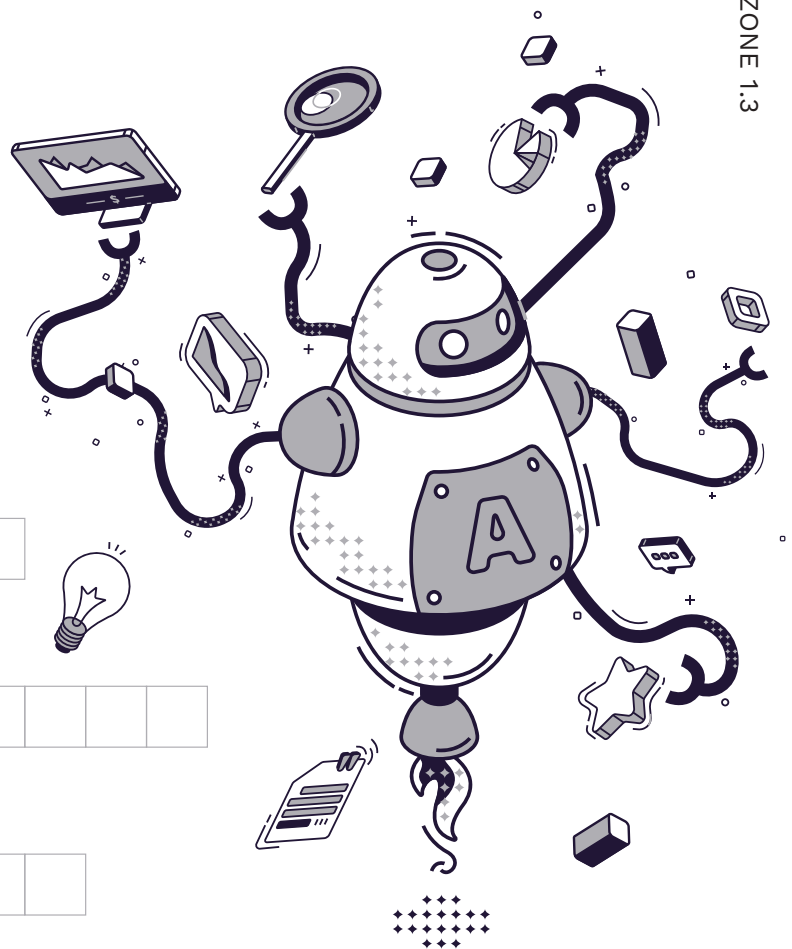
--	--	--	--	--	--	--	--	--	--	--

LCOVNOTUOIN

--	--	--	--	--	--	--	--	--	--	--

OAIHRGMITLC

--	--	--	--	--	--	--	--	--	--	--



Answers: Cluster, Training, Regression, Classification, Inference, Optimisation, Convolution, Algorithmic



| By Mohit Pandey

What Annoys Subbarao the Most?

“I open my LinkedIn and everyone is talking about LLMs and touting responsible AI as the focus point.”

T

The AI field is one crowded place these days, with almost everyone donning the hat of an ‘AI expert’. In order to separate the wheat from the chaff and get some clarity, AIM got in touch with Subbarao Kambhampati, professor of AI at Arizona State University for the past 32 years, and asked him what are the things that he loves, and what gets his goat.

Kambhampati’s views on LLMs align with Meta AI chief Yann LeCun. “We both feel that there is no reason to believe that these n-gram models would be able to reason beyond what they actually do,” he said. “They are just not made and meant to do that.” He gave an example of how LLM-based chatbots are still very weak at maths problems.

This brought up Q*, the OpenAI’s leaked model that was able to solve unseen maths problems. “It has stored an entire world in itself and is just able to retrieve information, that’s all. That’s pretty much the way to use LLMs, but that is a very dumb way to learn multiplication for humans; it’s the same for AI models.”

Talking about Perplexity AI, Kambhampati says that these are the classic cases of misunderstanding what LLMs are. “They are just retrieving information and summarising it, but they still hallucinate.”

He is interested in working with LLMs where they become the guessers. Kambhampati calls this LLM modulo fashion. Similar to LeCun’s vision of autonomous machine intelligence, he also proposes a world model, with different models connected together to a larger model, for bringing out information. “LLMs can only guess, they cannot verify.”



“Biology has become the refuge for people who want to do science but do not know maths,” he laughed, giving an analogy for people who talk about responsible AI, with the knowledge of neither AI, nor ethics.”

Kambhampati's views on LLMs do not end here. “Another thing that has changed is that instead of calling hallucinations a bug, people have started calling it a feature,” he added that most people are just getting impressed with what LLMs can do, and forget that the most important thing is to focus on accuracy, instead of just creativity, giving the example of calculators – “we want them to be perfect”.

So-called “AI Experts”

The problem that Kambhampati pointed out is that everybody thinks that they are an AI expert as the barrier to entry to become an AI ethicist, or voice for responsible

AI has been lowered. “I open my LinkedIn and everyone is talking about LLMs and touting responsible AI as the focus point,” Kambhampati laughed.

“Many people have come into AI only after 2013 – and some after ChatGPT – and often end up rediscovering the things that are already discovered.” “Biology has become the refuge for people who want to do science but do not know maths,” he laughed, giving an analogy for people who talk about responsible AI, with the knowledge of neither AI, nor ethics.

He also highlighted a conference that had 40 people discussing AI, with only one person who had actually written a research paper in their entire lives. “That’s definitely not just annoying, but worrisome,” he added that X and LinkedIn are just filled with cheerleaders of LLMs.

Recalling Mark Twain’s quote, “A classic is something that everybody wants to have read and nobody wants to read,” Kambhampati said that AI is the new classic as everybody wishes to know it, but they don’t really have the background to talk about it. “I don’t think you can get away with that just by speaking with confidence.”

He emphasised that people should learn the full spectrum of AI techniques and then use what you know is the right way to combine them. “But the problem is people do not even take an introductory course to AI, they just start working on deep learning projects.”

He also feels that education in AI is important, but only the basics are enough. “You can overdo AI education. You don’t need it, just like you don’t need to teach about electricity to everyone, but everyone uses it ...using calculators is different from understanding how it works.”

Not a fan of “UN of AI”

“As much as I hate the analogy, AI is becoming like electricity. But we do not have a global electrical policy. We do not need that for AI as well,” Kambhampati said. Moreover, he is also not a fan of making one organisation narrate all the rules around AI, as it would be ineffective. “Countries should be allowed to experiment with AI for their own use cases,” highlighting that advisory is acceptable, but it should be a mix of technology and policy experts.

Being passionate about open source, he said that the rise of Llama and Mistral, and then the rise of Indic LLMs such as Kannada and Tamil, has taken the baton away from OpenAI, which is very important. He highlighted that a lot of open source smaller models are performing better than models like GPT-4 that are much bigger in size on specific tasks.

“I hate the word foundational model,” Kambhampati added. “Just because there are different modalities coming



up such as videos and images, that does not mean that LLMs are starting to reason better.”

When it comes to India, Kambhampati points out that the problem with Indic LLMs is that there is more data from exam papers such as JEE and UPSC. “You can train an LLM to pass JEE, but not write something coherent in Indic languages,” he added, pointing out that this problem exists because most of the data in the Indic language is not yet digitised.

Most of the training data available today is in English, not in other languages, which Kambhampati believes is a very big problem for building large foundational models. “I said this in 2018, that data capture and digitisation would be a very big issue, but no one wants to take them up as they are unsexy.”

The Real Star of AI

“I kid around with my students that I was doing AI before it was cool,” said the self-proclaimed stand-up comic for his students.

Born in Peddapuram, Andhra Pradesh, Kambhampati went on to do his bachelor’s in electrical engineering at IIT Madras, and did his project on speech recognition that eventually got him into AI. “People would be sad that I am doing AI, when there are more exciting things like databases.”

Kambhampati admitted that his interest in LLMs started with all the hype around LLMs, as he believes that they are not just n-gram models, but can also do reasoning and planning. “But the claims that LLMs can do everything that an AI system should do, and it is complete, is not actually the case,” he explained. “They can do planning and reasoning, but not idea generation from scratch.”

Kambhampati is currently interested in working with the human-AI interaction field, which deals with AI systems that should be able to explain the decisions

to humans, instead of “silly saliency regions approaches where humans try to make sense of what machines are doing.” 🤖



“I hate the word foundational model,” Kambhampati added. “Just because there are different modalities coming up such as videos and images, that does not mean that LLMs are starting to reason better.”

| By Vidyashree Srinivas

If AI Works in India, it Can Work Anywhere

In the next 5-10 years, we would completely stop using the thumb, and use only voice-based interfaces,” the Wadhvani AI CEO stated.

Globally, India has the second-largest number of residents who use AI (81%). Further, Asian countries tend to display greater subjective knowledge of AI, with South Korea, China, and India leading the way. “India offers one of the most relevant and exciting contexts for AI, considering the readiness within the country to adopt various technologies, including IoT and AI, alongside other foundational technologies,” said Shekar Sivasubramanian, the CEO of Wadhvani AI, in an exclusive interview with AIM. Further, he said that the country’s ability to skip extensive wired network development and move straight to wireless infrastructure demonstrates its readiness and adaptability.

Sivasubramanian also highlighted that for AI to succeed on a large scale, access to diverse data, which is abundantly available in India, is required.

While we are at the early stages of harnessing this data, the country’s multilingual and multicultural environment provides a unique opportunity to develop context-centric, powerful AI solutions. “If AI works in India, it can work anywhere. This is because the diversity within the population acts as a natural protection against bias, allowing for more accurate and representative AI models,” added Sivasubramanian.

Sivasubramanian is an industry leader who brings 40 years of global applied technology and management experience towards creating positive and sustainable

Shekar Sivasubramanian
CEO, Wadhvani AI

impact at scale. He is currently driving non-profit organisation Wadhvani AI's efforts toward establishing AI-driven solutions and ecosystems for the benefit of millions across the developing world.

Voice is the Future

"In the next 5-10 years, we would completely stop using the thumb and use only voice-based interfaces," the Wadhvani AI CEO predicted.

He noted that to build meaningful voice-based interfaces, embracing the diversity of expression in India is crucial. And India's varied linguistic and cultural landscape provides a unique testing ground for developing advanced voice AI technologies.

Along similar lines, Pramod Varma, former chief architect of Aadhaar, told AIM, "Indian entrepreneurs should really look at voice as a completely new human-computer interaction method. It could be very powerful, and I think it's going to happen because voice is natural to humans."

In a previous interaction with AIM, Sarvam AI also mentioned that it is currently working on a voice-based Indic LLM, which it plans to release this year.

Earlier, chief AI scientist at Meta, Yann LeCun had said that in the next 10-15 years we won't have smartphones, and will be using augmented reality glasses and bracelets to interact with intelligent assistants. "The last thing we might want is intelligent virtual assistants that help us in our daily lives. So today, all of us here are carrying a smartphone in our pockets; 10 years from now or 15 years from now, we're not going to have smartphones anymore. We're going to have augmented reality glasses," said LeCun.

Even Meta CEO Mark Zuckerberg has repeatedly stressed that neural interfaces represent the inevitable next step beyond current methods like typing on screens.

Wadhvani AI Initiatives

India grows 26% of the world's cotton and nearly 100 million farmers rely on cotton farming for their livelihood. However, cotton is highly vulnerable to pests, causing yield uncertainty and financial distress to farmers.

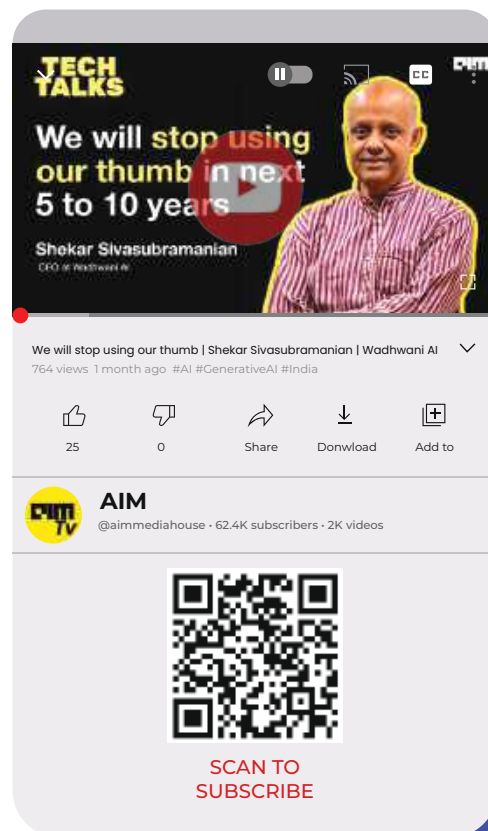
So, to help farmers protect their crops, Wadhvani AI has introduced CottonAce, an AI-powered early warning system available as an Android app.

Lead farmers, who work with welfare programs, use the app to upload photos of pests and the AI algorithm analyses the photos, determines infestation levels, and provides actionable advice, which is then shared with neighbouring farmers, even those without smartphones.

The app is available in nine languages, including English, Hindi, Marathi, Gujarati, Telugu, Kannada, Tamil, Odia, and Punjabi.

In the health industry, the non-profit institute developing AI solutions for social good is an official AI partner of the Central TB Division (CTD), and are developing multiple interventions across the TB care cascade and helping India's National TB Elimination Programme become AI-ready.

They use AI to interpret the results of the LPA test to determine drug resistance to TB. Each LPA strip encodes the drug-resistance pattern of the patient via a series of activated (dark) and inactivated (light) bands corresponding to different regions of the genome of the Tuberculosis bacterium. Further, it is developing multiple AI solutions to reduce morbidity and mortality for mothers and children in low-resource settings by improving the quality of primary care and strengthening the first 1,000 days of life. [com](#)



SHAPE THE
FUTURE OF AI

AIM
TV

AIM Videos includes
exclusive interviews with
CEO, CTO and **Founders** of
top companies.



Find tech explainer and latest
developments in artificial
intelligence by subscribing
AIM Youtube channel.

Scan to watch



ANGELS & DEMONS OF AI

In an exclusive interaction with AIM, Yann LeCun elaborated on his latest research and critiqued the chain of citations.

T
▼

The AI research community has made significant strides over the last few years, even as more is still being done. But, oftentimes, some of the work gets buried in the sand. Jürgen Schmidhuber seems to be overlooked. In July this year, he accused Yann LeCun of rehashing old ideas and presenting them without credit, saying—“We must stop crediting the wrong people for inventions made by others.”



“

I've known Jürgen since he was a student. He has an unusual idea about how credit should be attributed. It seems like he has the idea that if someone has the germ of an idea, they should get all the credit for what comes after. For most people, that's not how things work.”

The AI research community has made significant strides over the last few years, even as more is still being done. But, oftentimes, some of the work gets buried in the sand. Jürgen Schmidhuber seems to be overlooked. In July this year, he accused Yann LeCun of rehashing old ideas and presenting them without credit, saying—“We must stop crediting the wrong people for inventions made by others.”

In an exclusive interaction with AIM, Yann LeCun said: “I’ve known Jürgen since he was a student. He has an unusual idea about how credit should be attributed. It seems like he has the idea that if someone has the germ of an idea, they should get all the credit for what comes



after. For most people, that’s not how things work.”

Lecun (@ylecun)’s 2022 paper on Autonomous Machine Intelligence rehashes but doesn’t cite essential work of 1990–2015. We’ve already published his “main original contributions:” learning subgoals, predictable abstract representations, multiple time scales.

He believes that having an idea and experimenting with toy problems is not enough. “One can use these ideas at scale to make them work on real problems, which usually attracts attention. Then, one can deploy them in products. So, there is a whole chain of contributions,” added LeCun. “Sometimes, what’s hard is actually to instantiate those ideas into things that work. That’s where the difficult part starts very often. I can write ‘f’ of ‘x’ equals zero. Absolutely every theoretical statement in all of science can be reduced to this data. If you know what ‘f’ and ‘x’ mean, you might have a general idea, but then you need to contribute something concrete and operationalise this idea. He’s not the only one to have this attitude, but he’s a bit extreme,” quipped LeCun.

The drama unfolds

Schmidhuber started a whole dispute about recognition but that wasn’t a solitary instance. Earlier, news of the 2019 Honda Prize being awarded to Geoffrey Hinton “for his pioneering research in the field of deep learning in artificial intelligence (AI)” was expected to prompt the machine learning community to toast the man. Instead, the gloves came off, and an unexpected Internet dust-up ensued. Reason being: Schmidhuber’s 6,300-word blog post.

“ Humans do this all the time; we just don’t realise it. Being able to plan requires the ability to predict what is going to happen as a consequence of your actions. But AI can’t predict all the details. So, learning how the world can be affected by actions you take is an essential component of intelligence.”



“ Sometimes, what’s hard is actually to instantiate those ideas into things that work. That’s where the difficult part starts very often. I can write ‘f’ of ‘x’ equals zero. Absolutely every theoretical statement in all of science can be reduced to this data. If you know what ‘f’ and ‘x’ mean, you might have a general idea, but then you need to contribute something concrete and operationalise this idea. He’s not the only one to have this attitude, but he’s a bit extreme.”

At NIPS (now NeurIPS) in 2016, Schmidhuber interrupted a widely-attended workshop on GANs (Generative Adversarial Networks) by Ian Goodfellow. He claimed that he did very similar work earlier, which was overlooked. This caused outrage in the media, massive discussions on Reddit and Twitter, and certain famous people exchanged certain serious words. However, the community generally sided with Goodfellow. Simply because Schmidhuber “had the idea” doesn’t mean he is owed the credit for other peoples’ work decades later.

More to the story

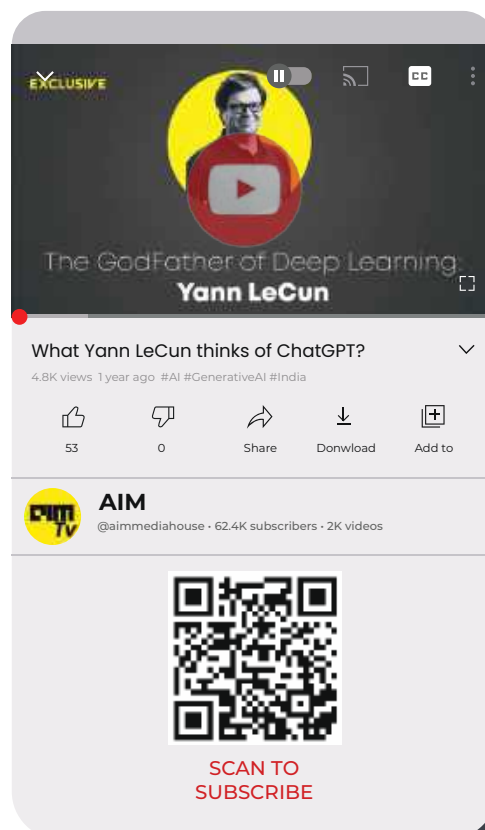
The tussle with LeCun was the research paper ‘A Path Towards Autonomous Machine Intelligence,’ which was published earlier this year.

Talking about the research paper, LeCun said that it focuses on integrating systems and understanding how the world works into a cognitive architecture. An architecture that allows machines not just to recognise, but also reason, in a hierarchical fashion. “Every human and a lot of animals can plan complex tasks by decomposing them into simpler ones. However, that seems to be something machines do not know how to do,” explained LeCun.

Further, he said, planning ability was a significant topic of interest in classical AI in the 1970s and 80s, but in the current brands of machine learning-based AI, there’s very little planning most of the time, except in certain types of AI systems for games like Alpha Go or chess where you need to plan. But, even with the latest system like Meta’s CICERO, there is quite a bit of planning there, but it’s not hierarchical planning.

Hierarchical planning is a problem-solving approach that focuses on decomposition, where problems are broken down step-wise into smaller ones until it is solved. “Humans do this all the time; we just don’t realise

it. Being able to plan requires the ability to predict what is going to happen as a consequence of your actions. But AI can’t predict all the details. So, learning how the world can be affected by actions you take is an essential component of intelligence,” concluded LeCun. 🇳🇵





THE MACHINECON USA 2024



SCAN FOR
MORE DETAILS

In its second edition, MachineCon USA was grander and more impactful than last year. The event was successfully concluded on July 26 at the historic 583 Park Avenue in New York City, with over 350 AI and analytics leaders showcasing their innovations, sharing unique solutions, networking, and more.

Additionally, AIM Research had the opportunity to honour thought leaders with the AI100 Award 2024, recognising and celebrating the achievements of individuals and organisations that have made significant advancements in the field of AI and analytics in enterprises. 🏆



AWS SUMMIT, WASHINGTON



AIM was instrumental in covering insights and in-depth analysis from the 2024 AWS Summit in Washington, DC—a pivotal event, with generative AI innovation and inspiration taking center stage. Dave Levy, vice president of Worldwide Public Sector at AWS, opened the keynote with a compelling video showcasing how AWS public sector customers are leveraging generative AI to create new possibilities.

Levy announced the AWS Worldwide Public Sector Generative AI Impact Initiative, a two-year, \$50 million commitment to accelerate generative AI innovation in public sector organizations. “This initiative is a significant step towards solving society’s most pressing challenges using AWS services and infrastructure,” said Levy.

Key announcements also included the AWS IMAGINE Grant: Children’s Health Innovation Award, a \$7 million grant aimed at advancing pediatric healthcare and research, and a \$3 million philanthropic commitment

to support nonprofit organizations focused on children’s health.

The keynote featured insightful talks from notable guest speakers. Lakshmi Raman, the CIA’s Director of Artificial Intelligence Innovation, discussed the agency’s approach to AI and the importance of workforce development.

Young Bang, the U.S. Army’s Principal Deputy Assistant Secretary for Acquisition, Logistics, and Technology, shared insights on the Army’s AI strategy and the potential benefits of AI in defense.

Dr. Adam Resnick from the Children’s Hospital of Philadelphia highlighted how cloud-enabled generative AI is accelerating breakthroughs in pediatric health.

The summit also underscored AWS’s commitment to sustainability, with a new study revealing that AWS’s infrastructure is up to 4.1 times more efficient than on-premises solutions. 🌱

Inside World's Largest Data & AI Gathering Ever

AIM was at the forefront, providing comprehensive coverage of World's largest data and AI gathering ever, The Data + AI Summit 2024.

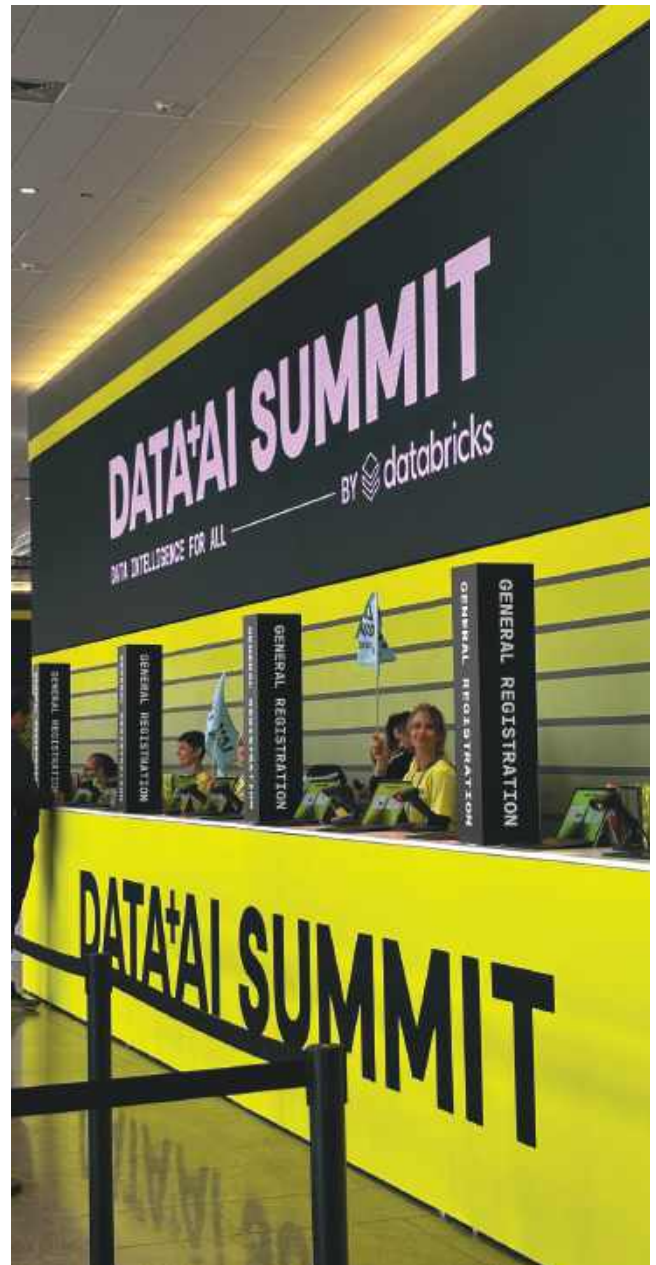
This landmark event witnessed over 16,000 in-person attendees and more than 40,000 virtual participants. Spanning four days, the summit featured major product announcements, extensive training sessions, certifications, partner showcases, and hundreds of breakout sessions. This gathering of the Data and AI community provided a valuable platform for sharing insights, networking, celebrating achievements, and collaborating on the future of these technologies.

Key sessions included insights from over 1,000 tech executives at the Executive Forum, where Databricks' co-founder and SVP of Field Engineering, Arsalan Tavakoli, emphasised, "Compound AI is pivotal in operationalising Gen AI, and open formats and standards are crucial for modern architectures."

Highlights of the summit also included significant themes such as the pressing demand for GenAI, with organizations eager to implement it in production to ensure quality AI responses without hallucinations. "Every organization wants to implement Gen AI immediately in production," noted the summit's key findings. Additionally, the growing challenges of security and privacy were discussed, alongside the fragmentation of data estates.

Databricks introduced several innovative products, including Mosaic AI, Unity Catalog Open Source, and Databricks AI/BI.

We are looking forward to next year's event, scheduled for June 9-12, 2025, which promises even more groundbreaking insights and advancements in data and AI. 🚀



10 INFLUENTIAL AI POSTS





Clem Delangue

CEO & Co-Founder
Hugging Face

Six predictions for AI in 2024:

A hyped AI company will go bankrupt
or get acquired for a ridiculously low
price

Open-source LLMs will reach the level
of the best closed-source LLMs

Big breakthroughs in AI for video, time-
series, biology and chemistry

We will talk much more about the cost
(monetary and environmental) of AI

A popular media will be mostly AI-
generated

10 millions AI builders on Hugging Face
leading to no increase of unemployment



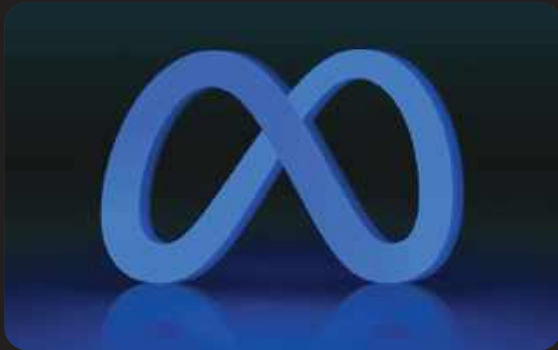


Yann LeCun

Director of AI Research
at Facebook

Do not work on LLMs

“If you are a student interested in building the next generation of AI systems, don’t work on LLMs. This is in the hands of large companies, there’s nothing you can bring to the table.”



Pedro Domingos

Professor of computer
science at UW

“AI that costs \$20/month is very, very far from AGI.”

“Anthropic is the most overrated company in AI.”



Bojan Tunguz

Machine Learning ex
Nvidia

“OK, now that I’m out, I can finally say this publicly: LOL, no, sorry, you are not catching up to NVIDIA any time this decade.”



**Andrej Karpathy**

Director of AI @ Tesla

“The killer app of LLMs is Scarlett Johansson. You all thought it was math or something.”

**Reid Hoffman**

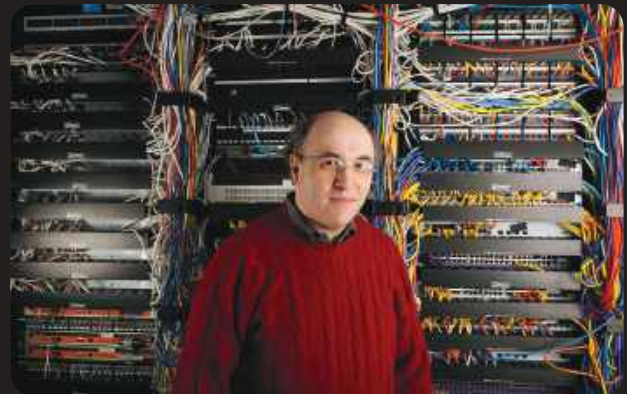
Co-Founder Reid Hoffman

“Traditional 9 to 5 jobs will vanish by 2034. “You may not only work at different companies, you might work in different industries with the gig economy revolution.”

**Stephen Wolfram**

Founder of Wolfram Alpha

“Aristotle founded or discovered logic by observing the world. ChatGPT thinks logically. Why? Because it notices all the logic in the data in its training set.”





Jonathan Ross

CEO of Groq

“We keep thinking of each technology as displacing work. One of the things that’s probably going to happen is we will create more jobs for people than we have people.”



François Chollet

Senior Staff Software Engineer at Google.

“There will be more software engineers (the kind that write code, e.g. Python, C or JavaScript code) in five years than there are today. The estimated number of professional software engineers today is 26 million, which would jump to 30-35 million in five years.”



Marc Andreessen

Co-author of Mosaic

“Do you think it’s finally possible for a one person billion dollar startup? You could put a whole bunch of things in this bucket...it’s the inherent scalability of software and the internet but you could also put AI.”



Empowering Leaders and Startups in the AI Era

World's biggest community exclusively for senior executives in Artificial Intelligence



Close-Knit Community

Join a supportive network of AI founders committed to shared growth and collaboration.



Expand Your Network

Connect with industry leaders to unlock new opportunities and partnerships.



Brand Visibility

Elevate your brand through high-profile platforms and industry recognition.

Unlock exclusive access to the AIM Leadership Council, where influential industry leaders come together. You're invited to join this esteemed group of professionals.



info@aimresearch.co

Yoshua Bengio
Computer Scientist





| By AIM Editorial

THINK LIKE YOSHUA

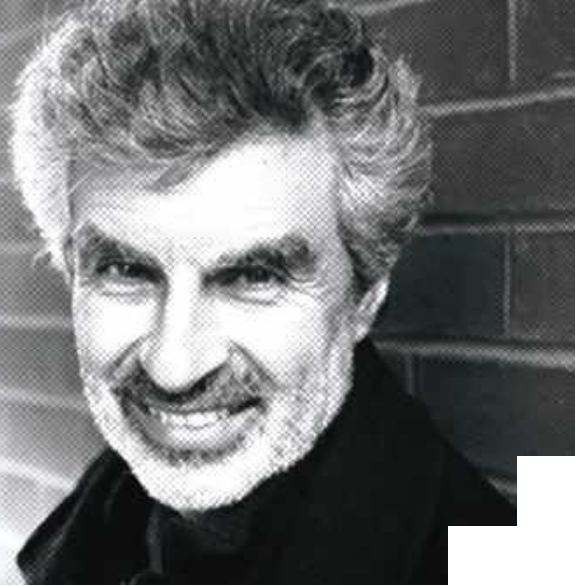
In an exclusive interaction with AIM, Yoshua Bengio takes us through the three decade memory lane.



I got very excited, and I still am. But, if you're doing work in an area that you don't feel a strong, emotional appeal to, then it's not going to be the same as if you're passionate about it."

“

There's an incorrect view of how science happens that you see in movies, where scientists discover amazing things alone in their garage or hidden mountain. It's not how science works. Science is about building on other people's ideas and doing it in groups.”



Widely recognised as one of the founding fathers of Deep Learning, Yoshua Bengio has played a central role in transforming artificial intelligence. He tutored a bevy of geniuses, including Ian Goodfellow. Bengio won the fabled Turing Award along with Yann LeCun and Geoffrey Hinton, the trio best known for tapping into the brainpower of machines.

AIM caught up with Yoshua Bengio to learn about the revolutionary's journey in the field and his learnings spanning over three decades.

Bengio considers his motivation to understand how things work and ask the 'why' questions an important factor that has kept him going. At some point, at the beginning of his grad studies, Bengio stumbled upon connectionist models. The term was used for the early work on neural networks—inspired by cognitive science and neuroscience—and applied to build learning machines, much like the work of Geoffrey Hinton.

Hinton was praised for his work on artificial neural nets, especially how they can be designed to learn without the aid of a human teacher. It was considered the start of autonomous intelligent brain-like machines.

Science is about building on other people's ideas

Interested in the notion of intelligence early on, young Bengio thought that researchers could understand human and animal intelligence to build intelligent machines, something he hadn't considered possible. "I got very excited, and I still am. But, if you're doing work in an area that you don't feel a strong, emotional appeal to,

then it's not going to be the same as if you're passionate about it," he said.

Moving forward, he emphasised the importance of collaboration. "There's an incorrect view of how science happens that you see in movies, where scientists discover amazing things alone in their garage or hidden mountain. It's not how science works. Science is about building on other people's ideas and doing it in groups," he said.

One of Bengio's most lauded works is with computer scientist and Google Brain researcher Ian Goodfellow on generative adversarial networks (GANs), which has revolutionised the fields of machine translation, image generation, audio synthesis, and text-to-speech systems.

Learning about each others' insights and brainstorming sessions with other scientists as well as his students is something Bengio considers very powerful. "If I was doing my thing alone, I don't think I would have achieved what I have. So, it's not only like an intellectual thing, but when we're together trying to understand or figure out something, we are motivated by the energy of others. We want to find solutions to problems. Especially when you're younger, all of these human social factors of the collaboration really make a difference," he added.

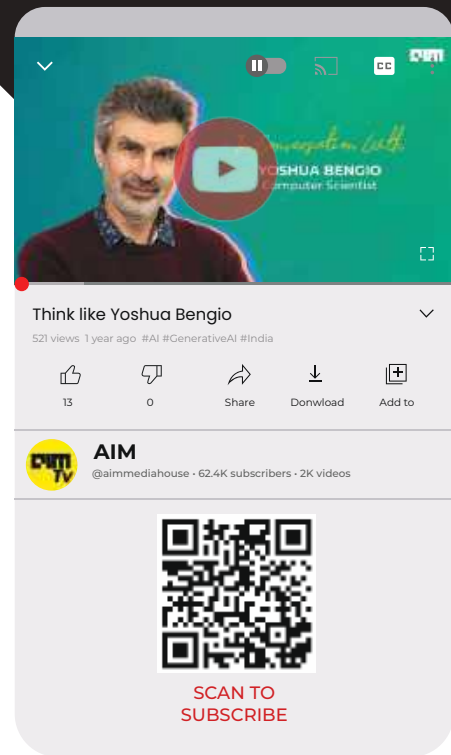
Talking about his initial days at work in retrospective, Bengio said there was a period when he worked with a not very popular thing in the science community, i.e., neural networks. He had a hard time convincing his students to continue working on the topic but decided to follow his intuition. As a result, Bengio was one of the first to combine neural networks with probabilistic models of sequences—a concept that has been extended to contemporary speech recognition systems.

In a paper published nearly two decades ago, he introduced the concept of word embeddings, a language modelling and feature learning paradigm in which words or phrases from a vocabulary are mapped to vectors of real numbers.

Research is an exploration

“The lesson is that people tend to do the same thing as others, and we need people to look in different directions. This is how we can be the most efficient. Research is an exploration,” Bengio said. Further, he explained that there’s exploration and there’s wasting time; the line between the two is not obvious. One needs to reason and be rigorous instead of being fooled by delusional ideas. Rationality is extremely important to fight the instinct to go with the flow in order to explore new territories.

Lastly, Bengio values the sense of humility. He said people who are socially recognised are often blinded by ego. “It makes us think we are right when we run out of sense. Another thing which I’m 100% about is that you have to prove and renew every day. So, that sense of humility is something that needs to be cultivated.” 🇨🇦



Ashan Willy
Chief Executive Officer, New Relic



Observability Tools Now Monitors LLMs

New Relic AI can now monitor AI-specific metrics like token usage, costs, response quality, as well as integration with their traditional application performance monitoring.

By **Pritam Bordoloi**

For cloud-native businesses, observability platforms are essential for gaining insights into application performance. DevOps practices thrive on the foundation of observability, empowering teams to iterate rapidly and deliver high-quality applications.

The Indian startup ecosystem is rich with cloud-native businesses. New Relic, one of the leading observability platforms, serves close to 12,000 customers in India. “I’ve visited India multiple times, and the fact that you can get things delivered in under 10 minutes is unparalleled. Trust me, I’ve been to many places and this does not happen anywhere. “Therefore, the demand for the applications and performance skyrockets here. A lot of these companies – the Swiggies and Ola Cabs of the world – use our observability platform,” Ashan Willy, CEO at New Relic, told AIM in an exclusive interview.

New Relic, with over 80,000 customers worldwide, caters to startups as well as large enterprises. Willy believes AI will bring a transformational change in the observability space. “I think at every inflection point, observability took a step forward and it will be the same with AI,” he said.

LLM Observability is here

New Relic’s comprehensive platform offers over 30 capabilities, delivering a seamless and connected experience throughout the various layers of the technology stack at every stage of the software lifecycle.

The company is now adding generative AI to the mix.

Last year, the company introduced New Relic Grok, which, according to them, is the world’s first generative AI assistant designed specifically for observability.

Willy believes, most companies today leverage LLMs in some way, and hence, “Observability offers the chance to monitor applications comprehensively, providing insights into both the application’s end-to-end performance as well as the LLMs”.

For cloud-native businesses, LLMs are often integrated with their existing operations. New Relic AI monitors AI-specific metrics like token usage, costs, response quality, as well as integration with their traditional application performance monitoring.

Other players in the space like DataDog and Dynatrace have also introduced LLM observability solutions as part of their platform.

Moreover, generative AI will enhance predictive analysis and reduce false positives, leading to more accurate and actionable insights. “One common challenge with AI/ML is false positives, leading to unwanted alerts and actions based on inaccurate information. However,



The Cost of AI

“

New Relic, with over 80,000 customers worldwide, caters to startups as well as large enterprises. Willy believes AI will bring a transformational change in the observability space. “I think at every inflection point, observability took a step forward and it will be the same with AI.”

with generative AI providing contextual insights, predictive analysis becomes more feasible. “Today, there is nobody who can really drive responses on top of the AI information yet but that’s going to change.”

AI is Democratising Observability

New Relic Grok allows users to ask questions about their systems in natural language and get insights. Willy believes this will democratise observability in a big way. “Why shouldn’t you have more of a Google-like UI where you come in and ask a question in plain English rather than learning some sort of crazy language, right? “So I think the first thing it does is democratise observability. Under the New Relic AI umbrella, we released something called Grok, an AI assistant before [Elon] Musk stole the name,” Willy laughed.

Moreover, AI also presents a huge opportunity for companies in this space to grow. “If there are roughly 30 million developers in the world, only about 2 million of them deal with observability on a daily basis. This presents us with a big opportunity,” he said. “How is Jensen Huang going to make his next trillion dollars? It’s not by selling chips to the data centre, the hyperscalers, because they are going to white label that. So, he’s got to move that stuff to the edge. You want to move it close to the edge which means there’s more things to observe,” Willy pointed out.

The Industry is Asking for a Consumption Pricing Model

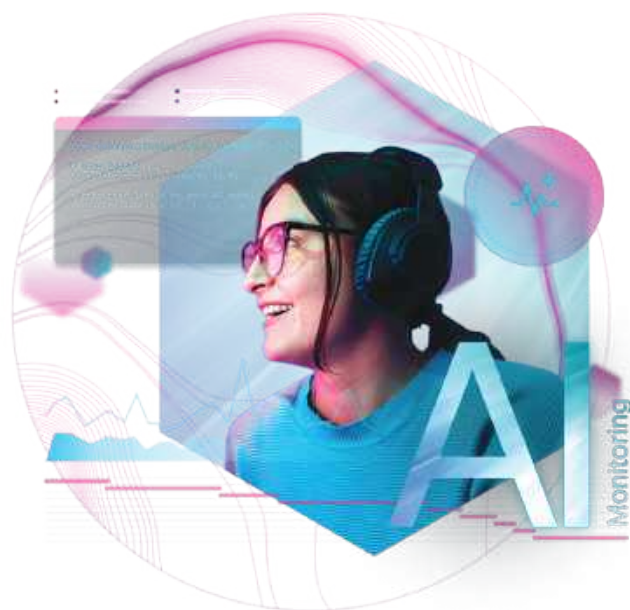
In recent years, as cloud costs rose, so did the expenses associated with observability, presenting a significant challenge for numerous cloud-native enterprises.

To address this, New Relic has introduced a price consumption or compute model, which means customers pay only for what they consume. “We’ve introduced things like telemetry data, pipelines, all of that, and cost management tools. I think that’s a big thing. So how do I observe everything I have in a way that makes it effective for me? “Hence, we have introduced the consumption-based pricing model which customers love, but the investors, not so much. But I think over time, if customers like it, investors will come to like it as well.”

Furthermore, Willy also points out that New Relic is the only company in the core observability space providing a consumption-based pricing model.

New Relic is Embracing Open Source

The domain of open source continues to hold sway, retaining relevance in various domains in contemporary times, including observability.



“How is Jensen Huang going to make his next trillion dollars? It’s not by selling chips to the data centre, the hyperscalers, because they are going to white label that. So, he’s got to move that stuff to the edge. You want to move it close to the edge which means there’s more things to observe.”

New Relic, too, is embracing open source. “A lot of our customers are asking us to embrace open source and open telemetry, which is a portion of open source that is very relevant for the observability space. We’re embracing that in a big way,” Willy said.

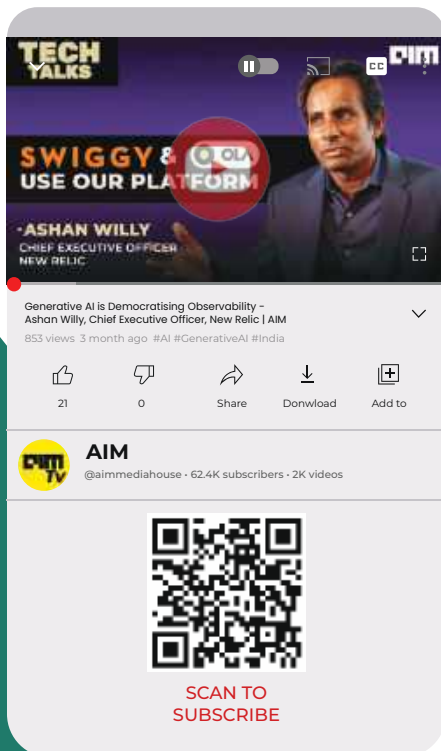
Modern cloud-native applications are distributed, making the capture and export of telemetry data complicated. Open Telemetry standardises telemetry data generation, collection, and exportation.

With vendor-neutral APIs, SDKs, and tools, this open-source framework enables organisations to instrument applications universally, promoting data delivery to diverse observability backends.

Willy’s statement does hold true as Relic is one of the major contributors to the open telemetry project. The company has open-sourced many of its agents, integrations and SDKs under open source licences. 



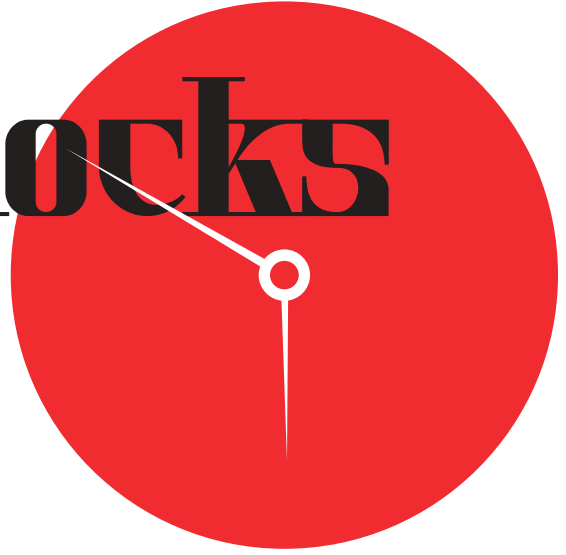
“A lot of our customers are asking us to embrace open source and open telemetry, which is a portion of open source that is very relevant for the observability space. We’re embracing that in a big way.”



By Vandana Nair

BAIDU CLOCKS 6,000 DRIVERLESS RIDES PER DAY IN WUHAN, CHINA | 94

Baidu Clocks 6,000 Driverless Rides Per Day in Wuhan, China





“

I think we are all at L4 today, and with the government regulations, it's not possible to do L5,” said Helen Pan, the GM of Baidu Apollo.



Baidu is clearly the darling of autonomous giants. A few months ago, Tesla chief Elon Musk visited China, the second largest market for the company, to release FSD (full-self driving) robotaxis in the region later this year.

However, this was made possible only through a partnership with Baidu, which is already slaying in the autonomous segment with its fleet of robotaxis making it the largest autonomous ride-hailing service in the world.

Baidu's robotaxis are currently available in four cities in China, including Wuhan, Beijing, Shanghai and

Chongqing. In Wuhan alone, it caters to a population of seven to eight million covering an area of 3000 sq km. "Here [Wuhan], we have 6000 driverless rides per day. So, all of them combined together is quite a big achievement," said Helen K Pan, general manager and board of directors for Baidu Apollo, California, in an exclusive interaction with AIM.

Though robotaxis currently operate in only four cities, their autonomous driving service is available in 11 cities across China.



“What I see in China is that the cost of the robot taxi is really low. We have to have the cost constraints much tighter than the rest of the world.”

In May, Waymo, a self-driving cab service, confirmed that they are operating more than 50,000 paid trips every week across three cities in the US.

Level 5, Not Yet

Baidu has been working in the driverless segment for over a decade and has developed various levels of autonomy. It currently operates at Level 4. However, their technology borders on Level 5 too. "Typically, we call it L4 and L5, with L4 [autonomy] implying in geo-fenced regions," said Pan. "With the government policy, we are actually more in the L4 region."

As of April 2024, the cumulative test mileage of Apollo L4 has exceeded 100 million kilometres. While Baidu has been advancing its autonomous capabilities for over a decade, the West is already competing in this space with Tesla and Waymo. However, Level 4 is what

they all are at. "I think we are all at L4 today; and with the government regulation, it's not possible to do L5. Another thing is that, I think, all of us providing this technology haven't been tested in all the scenarios. We wouldn't have this confidence to claim that we have the L5 capability."

Baidu vs the World?

Pan believes that the training conditions for Baidu's vehicles act as the biggest differentiator when compared to Tesla or Waymo. "China has a couple of unique market technology challenges compared to the US and other countries," said Pan. "The roads are in a diversified condition in China. The driving behaviour, the population density, and the driving scenarios are much more complex."

Pan also elaborated on how the aggressive driving

behaviour of drivers in the Chinese market, especially in certain regions of the country, contributes to building rich data. “Outside China, you’ll take a longer time to see, but you will face those corner cases here more frequently. So, that’s a benefit,” she said.

In addition to rich data, costing is another differentiator. “What I see in China is that the cost of the robot taxi is really low. We have to have the cost constraints much tighter than the rest of the world.”

Interestingly, Pan comes with a rich experience of working with autonomous technology. Before Baidu, she worked with Google and Waymo on their self-driving technology.

Challenges Remain

Though comfortably placed in the autonomous vehicle (AV) space, Baidu is still experimenting to navigate the challenges associated with it. “If you look at it from an AI point of view, I think it’s challenging. AI algorithms are evolving,” said Pan, who explains that you need real-time and extremely high-accuracy data.

Pan confirmed that the company is also looking at developments around large models. “We’re providing this development, an end-to-end large model and we are able to tailor all the diversified scenarios. “It will also help us migrate from one city to the next, once you have this large model instead of a very customised one for a particular city. So, I think that’s quite a big breakthrough for us to be able to further expand our technology into multiple regions in China,” said Pan.

Safety Regulation

Pan discussed the government’s apprehensions associated with AV technology.

“The government actually wants to embrace this new disruptive technology. But on the other hand, they don’t know how to really regulate it. They want to foster this technology, and make sure it’s not chaotic and that they can really control this one,” she said.

To address this, Baidu is actively working with the Chinese transportation department to help form regulations to ensure safe deployment. “We run a lot of simulations, and we also do a lot of early testing with the safety driver behind the wheel, and proving our technology is working. Then, we gradually remove the safety driver, and become completely driverless. So, when we reach the driverless stage, we know our technology, and we have the confidence,” said Pan.

Furthermore, there are 10 levels of redundancies built into the autonomous driving systems, and AI algorithms. The Chinese government also mandates specific test conditions and routes to obtain a driverless certification.

When asked if 100% safety can be achieved, Pan was

“What I see in China is that the cost of the robot taxi is really low. We have to have the cost constraints much tighter than the rest of the world.”



Robotaxis is the Future

- **Waymo**

As of 20th August 2024, Waymo clocked

100,000

driverless robotaxi rides per week.

- **Baidu**

Delivered 899K rides with over

7 million

total rides in July.

- **Tesla Robotaxi**

Anticipated to hit the market soon, with Elon Musk believing they’ll have an autonomous

fleet of 7 million that could reach **10-20**

million in the future.

pragmatic in her answer. “Nobody can say everything is 100% safe, but I think it will evolve. If you look at it 5-10 years from now, we’ll still continue to improve our safety.” 🇨🇳

THE LAST WORD

Integrating Generative AI and LLMs into Your Data

Integrating Generative AI and large language models (LLMs) into your data landscape represents a significant shift in harnessing artificial intelligence for data-driven solutions. These technologies offer unprecedented capabilities to process, understand, and generate human-like text, transforming business interactions with data. By integrating Generative AI and LLMs into data infrastructure, organisations can enhance analytics, automate tasks, and drive innovation across sectors.

Current Market Sentiment

Reflecting on the market landscape in Silicon Valley, Ramaswamy acknowledged the challenges startups face, particularly due to high interest rates. He highlighted a cautious optimism regarding the economy, with significant excitement around AI. However, he also noted that startups will need to demonstrate the business value of their actions within a limited window of 18 months. "Some of the excitement around AI and what that technology can do is make people believe everything is fine with startups... I would give it like utmost an 18-month window," said Ramaswamy.

Snowflake's Integration of AI

Snowflake's strategy involves integrating Generative AI and LLMs to enhance its data platform. The company has developed Cortex, a platform layer with semantic search capabilities, allowing for easier interaction with data. This integration aims to democratize data access, enabling users to create applications such as chatbots or copilot tools for generating SQL with ease. "We want to take these, embed them, and create

an application for some of your users... the next logical step as an accelerant into democratising access to data in Snowflake."

Ensuring Secure Data Practices

Snowflake emphasized secure data practices when leveraging internet resources for applications like ChatGPT. Ramaswamy explained that while training on diverse internet data is crucial for linguistic capabilities, fine-tuning on specific enterprise data ensures precision and relevance. "We are building semantic search into the core of Snowflake to serve as a frame-setting or retrieval system... we emphasize how you make them feel safe," he added.

Handling Diverse Data Formats

Addressing the challenges of diverse data formats, Snowflake ensures that data is cleansed, timely, and reliable. This foundational work is crucial for enabling AI applications. Snowflake's AI team emphasises high precision and recall in generating SQL and other outputs, ensuring robustness and reliability. "There is no AI strategy without a data strategy... we create systems like this, we emphasise how you make them feel safe," said Ramaswamy.

Balancing Accessibility and Proficiency

Snowflake aims to balance accessibility to data-driven technologies with necessary user proficiency. The platform provides tools and guidelines to ensure safe and effective use, while also highlighting the importance of context-setting for chatbots and other



Sridhar Ramaswamy, previously a Co-Founder at Neeva and now Chief Executive Officer at Snowflake, provides valuable insights into these technological advancements. With a rich history at Google, where he advanced to the role of SVP of Ads & Commerce, Ramaswamy brings extensive experience to Snowflake.

AI applications.

"Every chatbot will have an introduction where it will say... this context setting is going to be important," noted Ramaswamy.

Responsible Use of AI

Ramaswamy emphasised Snowflake's commitment to responsible AI use, particularly in business applications. By integrating AI into the core of Snowflake and maintaining an AI governance council, the company ensures safe, reliable, and valuable use of these technologies. "It's really like having our customers access data more easily and freely, but also more safely and in a reliable way," said Ramaswamy.

In conclusion, integrating Generative AI and LLMs into data platforms like Snowflake's offers transformative potential for businesses. By ensuring secure, precise, and responsible use, these technologies can drive significant innovation and value across industries. 

AIM



AIM publishes every day, and we believe in quality over quantity, honesty over spin. We offer a wide variety of branding and targeting options to make it easy for you to propagate your brand.



info@aimmediahouse.com

THINK

A
I

