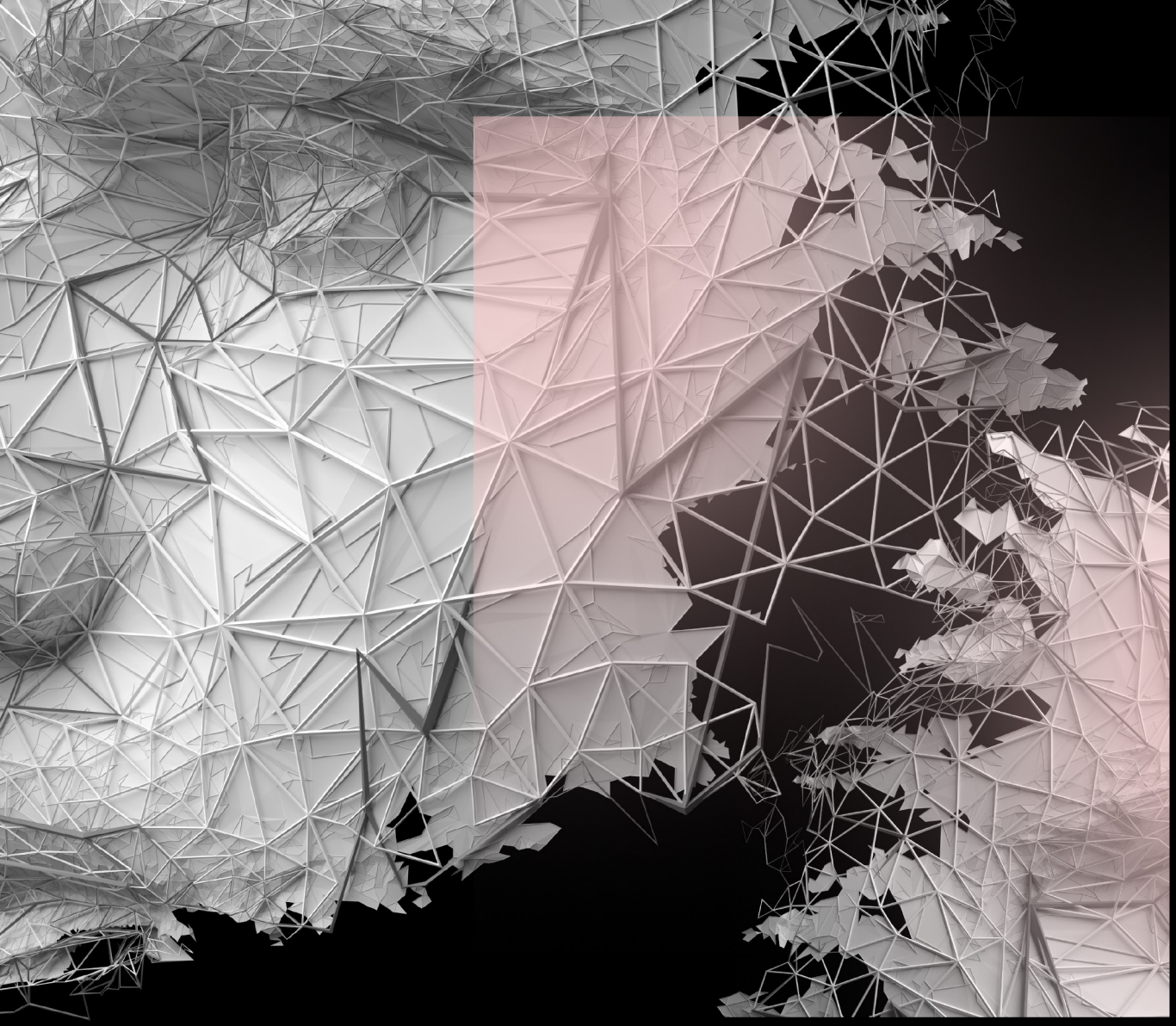


Sponsored by CLEARML

AI Infrastructure Alliance

Enterprise Generative AI Adoption

C-Level Key Considerations, Challenges, and Strategies for Unleashing AI at Scale

2023

When we started the AI Infrastructure Alliance in 2021, AI and machine learning was a niche topic, with a passionate and dedicated set of practitioners but it was mostly hidden from the general public. People used AI without thinking about it, like when they talked to their phone and it understood what they said or when they used Google Translate on their summer vacation.

All that changed over the last year with the release of **ChatGPT**, a revolutionary Large Language Model (LLM).

To say that ChatGPT was a titanic shift in public perception of AI is an understatement. It rocketed to over 100M users in just three months between February 2023 to April 2023. It hit its first million users in only 5 days. To put that in context, it had taken Netflix 3.5 years to hit 1M users.

Regular people everywhere were suddenly using AI directly, instead of hidden inside an app. The experience is raw, visceral and direct. Now everyone is aware of AI and has an opinion on it, whether it's governments rushing to pass landmark legislation, to artists either for or against AI, to big business, to the local cab driver at the airport.

Companies are racing to integrate generative AI and LLMs into their applications. But how are enterprises faring with weaving these models into their workflow and applications? Are they too unpredictable? Are companies having tremendous early success or really struggling? Or is the answer somewhere in the middle as enterprises learn how to work with these very new kinds of systems?

In this special report, the AIIA polled over 1000 businesses with over 1 billion USD in yearly revenue to see how they're using AI and whether they're successfully integrating LLMs and generative AI into their business and products.

Research and development is moving at a breakneck pace now, with new ideas, new research, new applications and models landing almost daily. Everywhere developers have raced to embrace the shift, building new kinds of apps on top of LLMs like GPT-4, the veteran LLaMA foundation model from Meta, Falcon, WizardLM, Starcoder and we've seen a flurry of state-of-the-art foundation models like SAM (Segment Anything Model), Stable Diffusion, Gen1 and Gen 2. The pace continues to speed up. But are enterprises able to keep up with the rate of change? Is anyone?

A few years ago, the general feeling in the MLOps industry was that everyone would have a huge team of data scientists and train advanced machine learning models from scratch. It looks like that future will never come to pass. That's because foundation models are hard to create, often costing millions, to tens of millions of dollars, to 100s of millions of dollars to train.
These models are also big. The largest transformer based foundation LLMs have memory requirements that scale quadratically with parameter count. Serving inference with LLMs often requires many datacenter level GPUs, hundreds of GBs of RAM, and backend tricks like weight streaming to make them work.

It can take hundreds or thousands of the most advanced GPUs running for months to train these models along with advanced supercomputing teams to manage them. Some analysts have noted that many of the top models lose money on inference as the world waits for economies of scale to kick in and drive down prices.

Because of all this, most companies are now turning to advanced foundational models created by a small subset of companies and researchers. They're looking to fine tune these base models to make them work for their personal use cases and needs and to get them integrated into their applications.

There's little doubt that since the release of ChatGPT, we've seen a Cambrian explosion of AI use cases and a massive uptick in public interest in AI.

But has that translated to Enterprise adoption and integration even as individual developers and small businesses adopt AI at a furious pace?

We set out to find out.

Demography

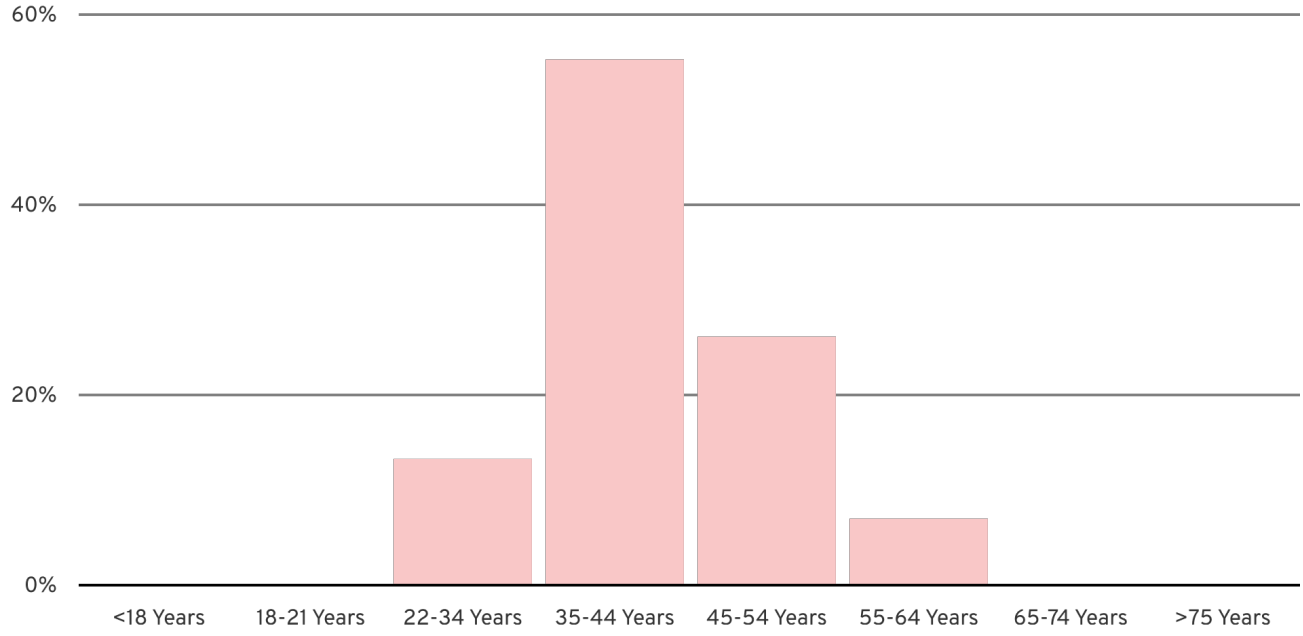
To start with, let's understand the demographics of who we talked with in our survey. Essentially, we talked to very big enterprises. There are no small or even medium businesses in our results. Every respondent had more than 10,000 employees and 50% had over 20,000.

We also primarily talked with leadership and the heads of teams, with job titles like CIO, CTO, Head of AI, VP of Data or VP of Artificial Intelligence. That means the results primarily represent the c suite and the team leads but not the engineers and their teams.

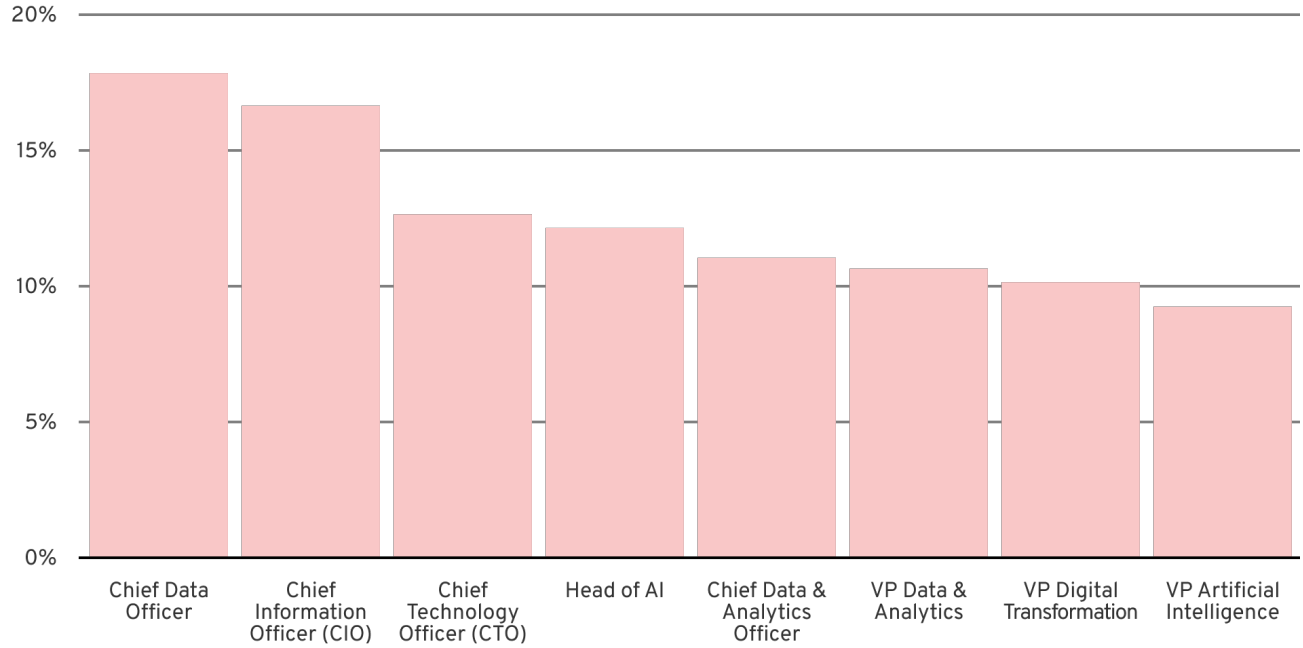
The survey primarily focused on companies in the US, UK and Canada, as well as across the EU. However, we did have a number of respondents from Japan and South Korea but we don't consider this survey a definitive representative of the Asian enterprise company perspective.

Lastly, we spoke to people across a large range of verticals, everything from law firms, to manufacturing, to telecommunications, energy, food, healthcare and more. The largest representations came from Information Technology companies, but no vertical surveyed represented more than 8% of the total respondents, so we had a wide range of views across a wide range of companies.

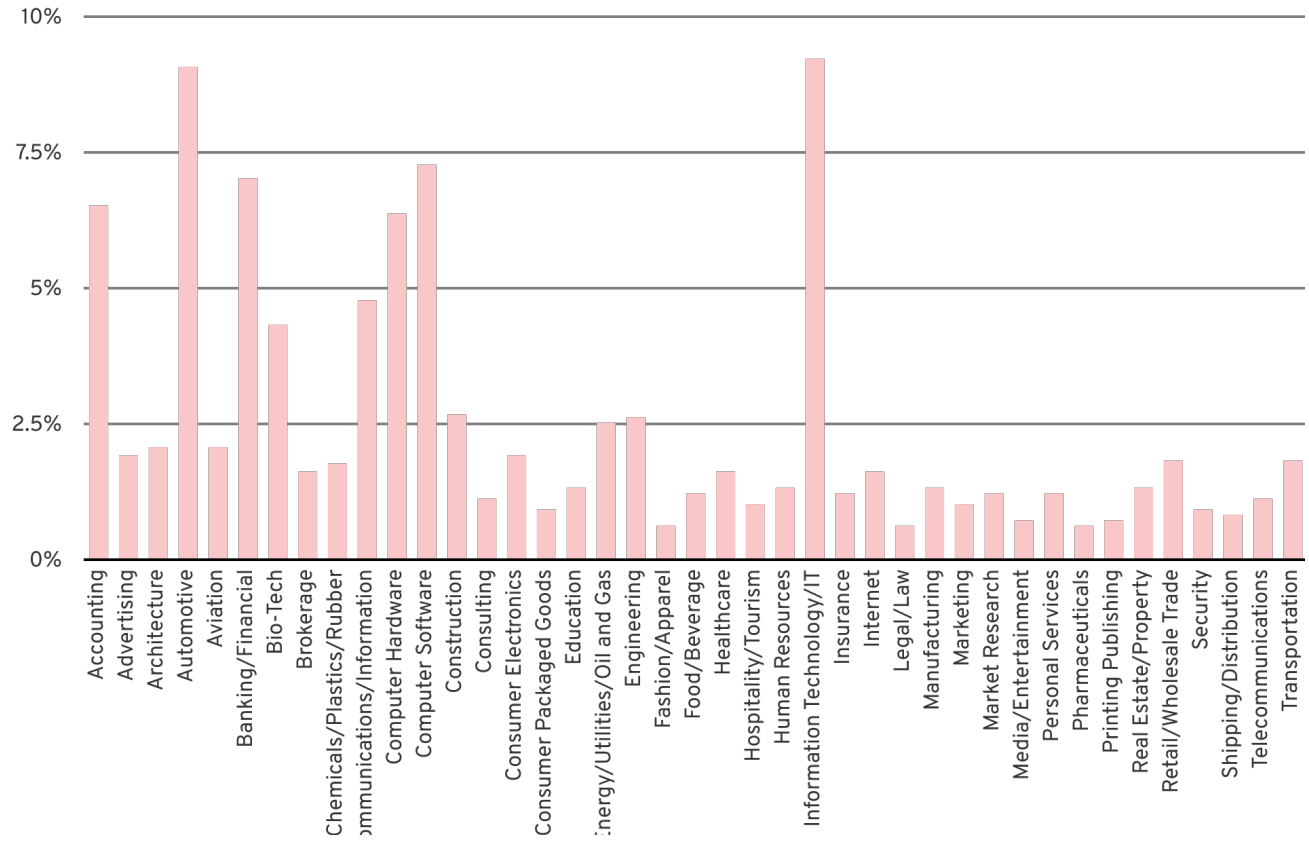
Age of Surveyed Respondents



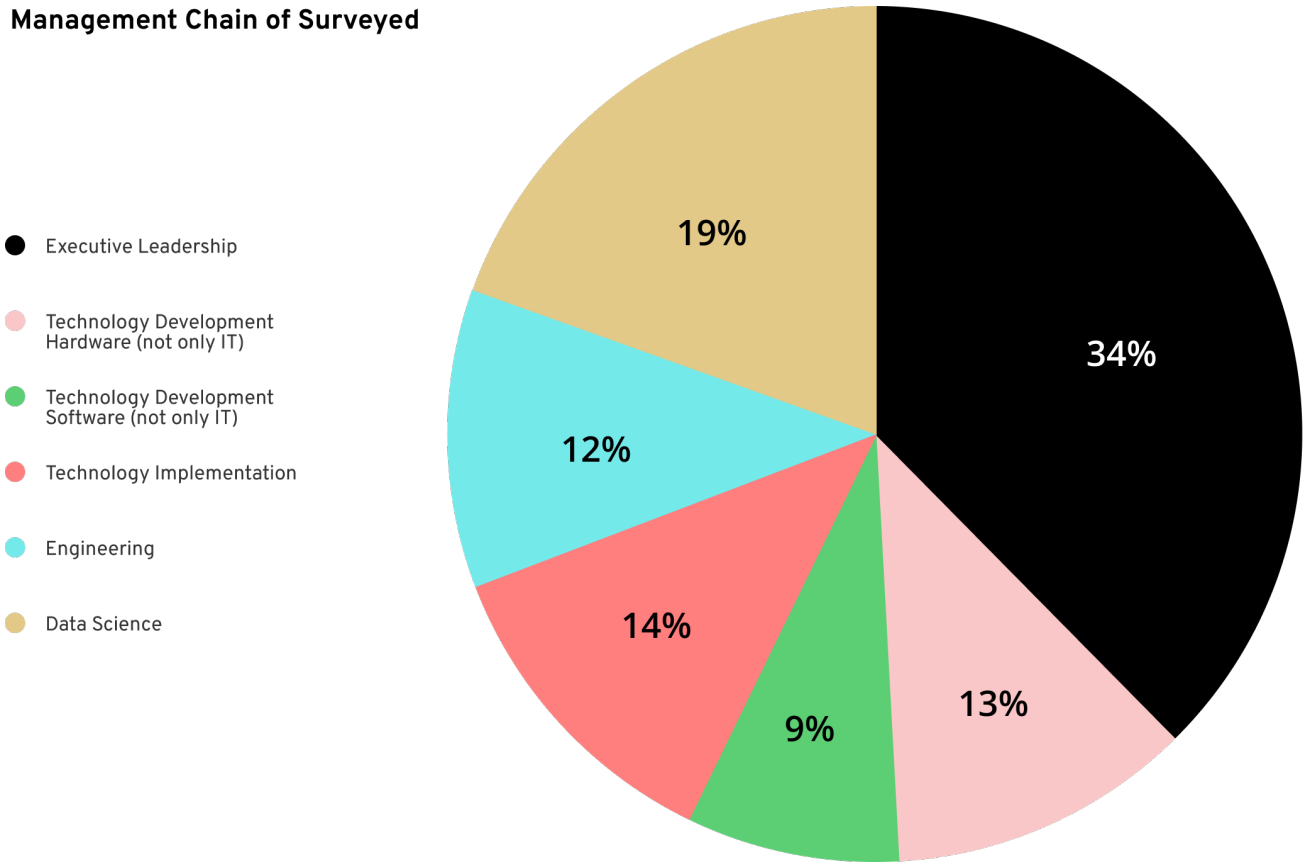
Title/Position of Surveyed Respondents



Industry Represented by Survey



Management Chain of Surveyed



The Big Finish

Despite some big losses from previous AI deployments and big challenges to getting AI working and running effectively in large organizations, nearly every company we surveyed was very enthusiastic about weaving LLMs and generative AI into their products and workflow. Not only that, they expect it to drive more revenue or reduce costs or both.

Maybe it seems strange that enterprises would remain upbeat and enthusiastic about AI after taking some early hits but it's likely a result of where we are in the development of AI. In the past few years, we were just starting to see where it could fit and what it could do at the application layer. We had a lot of data science and research but not as many applications. All of that is changing. LLMs are very general purpose tools and when you combine them with the power of other models for doing specific tasks or with external knowledge bases they promise to deliver the long hoped for value in AI. It's sometimes dangerous to say "this time is different" but it does seem that this time is different and that's why enterprises are lining up to bring these advanced capabilities in house as fast as possible.

But there's the real challenge. Speed. Even with all that enthusiasm, large enterprises face a much larger headwind of regulation and compliance that can be a headache for even the most well run teams. With the EU focusing on explainability and interpretability in models and applications, especially for applications deemed high risk, companies may find their enthusiasm only gets them so far as much of interpretability and explainability remains in the early stages of research. There are already advanced explainability systems for traditional models but as of yet there are no ways to effectively detect truly bad answers or logical flaws from LLMs. It's likely to take a lot more research to get there. At the AIIA, we expect regulators to take a gentler hand at the dawn of new legislation, while software catches up to the need for better understanding of what these systems are doing and how they make sense.

Yet even with all of these challenges, AI is weaving its way into more and more enterprise domains and companies remain optimistic that it can drive more revenue and slash costs. All of this points to one thing: We're entering an age of industrialized AI.

Enterprises will not be left out and will drive much of the shift. Industrialization is where we see the rapid acceleration of research into real world applications, supported by ever more advanced AI chips, better software, and ideas imported from other domains that cross pollinate with traditional machine learning. It marks a time when companies around the world pour huge amounts of money, people and time into the pursuit of ever smarter software and machines. Now regular people and traditional coders and business people are getting their hands on super powerful models and taking them in bold new directions. They'll bring with them their know-how from decades of traditional computing and make the models, stacks and pipelines safer, more resilient and more predictable.

AI has already busted out of the walls of Big Tech R&D labs. Now even companies like Google are back on their heels as newer, faster and more agile smaller companies sense a once in a generation chance to reset the order of things and build new tech powerhouses. This was unthinkable even a year ago but it's the reality now. AI is poised to disrupt the old business model of the web in major ways. While the incumbents have the early advantage, they also have innovator's dilemma. They need to protect their old business model which is centered on advertising. But what happens when someone no longer needs to go to that ugly recipe site filled with ads after every paragraph

because AI can just tell them how to make herb chicken and rice for dinner? That's when the game really starts to change and a new business model will have to emerge to replace the old one.

As enterprises and small businesses find more and more use cases for AI and drive more revenue, it will turbocharge development of AI, rapidly advancing and refining the ideas of the research labs. As more and more product people, technical people, and traditional coders work with super charged models they'll take us in directions none of the researchers could have possibly imagined. They'll make the models smaller and faster, and find ways to weave them together with other traditionally coded apps, all while figuring out better ways to put guardrails on them so they deliver more and more consistent results.

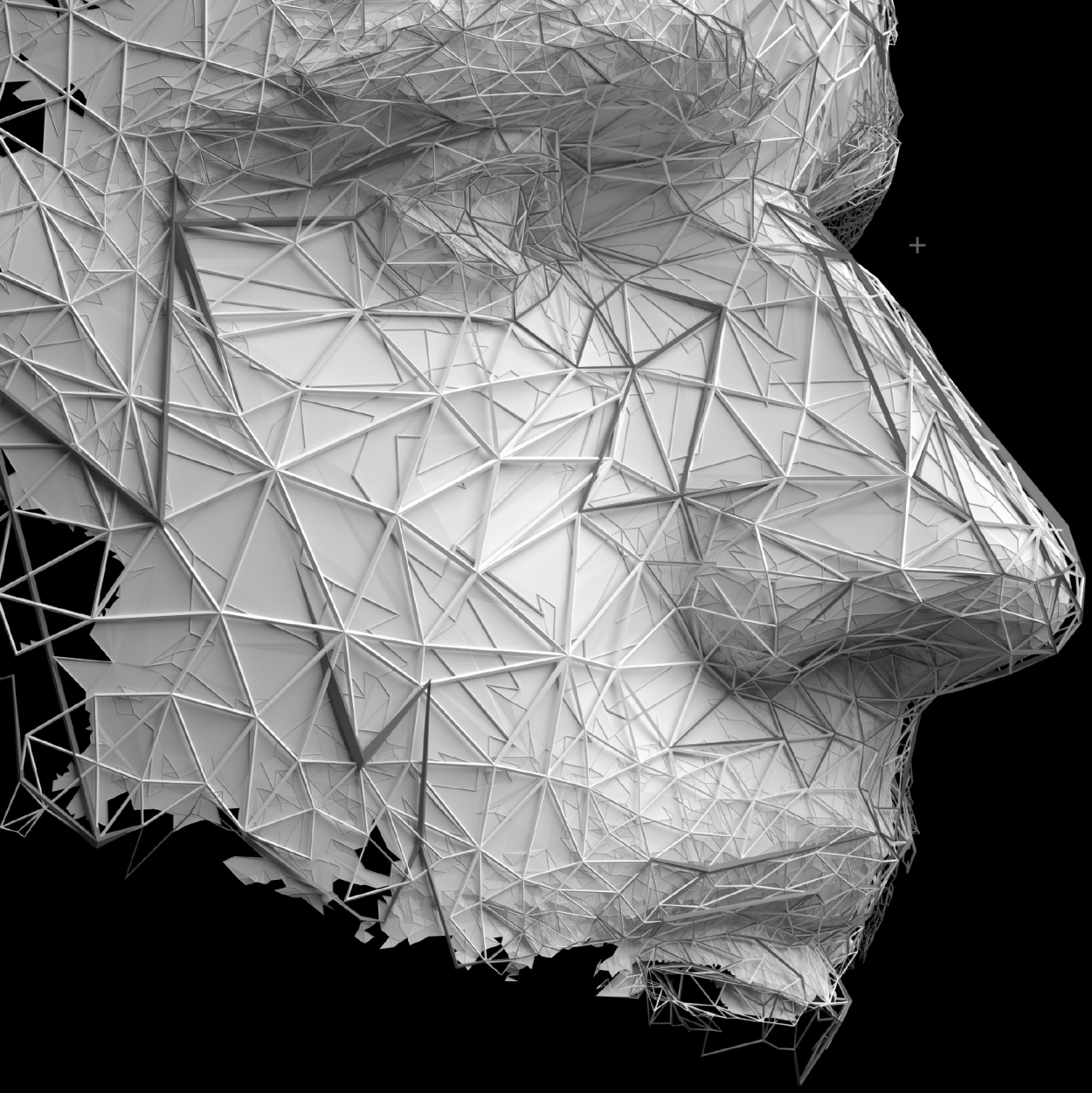
It's also industrialization that will push these systems to get safer, smarter and more aligned with what we want them to do. That's because companies and regulators and the general public will demand they get more reliable so they're more consistent, manageable and predictable. Not perfectly consistent because these are not and never will be deterministic systems, but they will get much more consistent, much more often.

We're also entering a time when many companies will never train their own models. They may fine tune one, but we expect even that to fall by the wayside as more and more models get stronger right out of the box. It's just too complicated and too expensive for the average company to build out a supercomputer, hire an ML team, a data team, an MLOps team and build a top of the line model. Expect more companies to look for finished products rather than train their own.

In the coming years, the vast majority of companies won't even interact with AI at a low level. That's like writing in Assembler, essential for a small subset of tasks, but way too complicated for most projects. Companies just don't have the time, money and people power to ingest billions of files, label them, do experiments, train a model, deploy it, optimize it and scale it.

But no matter what, big enterprises everywhere are already undergoing a remarkable transformation, weaving AI into every aspect of their business. We expect that to speed up in the coming years and to change many businesses dramatically as they drive untapped value from intelligent machines.

Enterprises are ready, willing and able to bring these systems in house and their efforts will make these systems better for everyone all up and down the economic chain. AI is poised to remake every aspect of the economy from logistics, to manufacturing, to healthcare and more. It's not a matter of if companies will overcome their challenges, it's a matter of when. It may not move as fast as everyone expects but make no mistake, big changes are coming to enterprises everywhere as AI weaves its way into every aspect of our lives.



AI Infrastructure Alliance

Website
ai-infrastructure.org

LinkedIn
linkedin.com/company/ai-infrastructure-alliance

Twitter
twitter.com/AiInfra

CLEAR|ML

Website
clear.ml

LinkedIn
linkedin.com/company/clearml

Twitter
twitter.com/clearmlapp